2017-07-01

Alternative Methods of Estimating the Degree of Uncertainty in Student Ratings of Teaching

Ala'a Mohammad Alsarhan
Brigham Young University

Alternative Methods of Estimating the Degree of Uncertainty in Student Ratings of Teaching

Ala'a Mohammad Alsarhan

A dissertation submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Richard R. Sudweeks, Chair
Joseph A. Olsen
Lane Fischer
Sterling C. Hilton
Ross Larsen

Department of Educational Inquiry, Measurement, and Evaluation

Brigham Young University

ABSTRACT

Alternative Methods of Estimating the Degree of Uncertainty
in Student Ratings of Teaching

Ala'a Mohammad Alsarhan
Educational Inquiry, Measurement and Evaluation, BYU
Doctor of Philosophy

This study used simulated results to evaluate four alternative methods of computing confidence intervals for class means in the context of student evaluations of teaching in a university setting. Because of the skewed and bounded nature of the ratings, the goal was to identify a procedure for constructing confidence intervals that would be asymmetric and not dependent upon normal curve theory. The four methods included (a) a logit transformation, (b) a resampling procedure, (c) a nonparametric, bias corrected accelerated Bootstrapping procedure, and (d) a Bayesian bootstrap procedure. The methods were compared against four criteria including (a) coverage probability, (b) coverage error, (c) average interval width, and (d) the lower and upper error probability.

The results of each method were also compared with a classical procedure for computing the confidence interval based on normal curve theory. In addition, Student evaluations of teaching effectiveness (SET) ratings from all courses taught during one semester at Brigham Young University were analyzed using multilevel generalizability theory to estimate variance components and to estimate the reliability of the class means as a function of the number of respondents in each class.

The results showed that the logit transformation procedure outperformed the alternative methods. The results also showed that the reliability of the class means exceeded .80 for classes averaging 15 respondents or more. The study demonstrates the need to routinely report a margin of error associated with the mean SET rating for each class and recommends that a confidence interval based on the logit transformation procedure be used for this purpose.

Keywords: student evaluations of teaching, confidence interval, reliability of class means, logit transformation, resampling, bias corrected accelerated, Bayesboot

ACKNOWLEDGEMENTS

First and foremost, all praise to Almighty Allah for the strength and guidance that enabled me to complete this work, for all the great people who have been the support that I need.

"الْحَمْدُ لِلَّهِ الَّذِي بِنِعْمَتِهِ تَتِمُّ الصَّالِحَاتُ"

I would like to express my indebtedness and thanks to my chair, Dr. Richard Sudweeks; you have been the remarkable mentor, teacher, and dear friend for me. Your advice on both research as well as on my career have been invaluable. Your enthusiasm, energy, and liveliness have inspired me to be the researcher I am today. I would like to thank my committee members, Dr. Joseph Olsen, Dr. Lane Fischer, Dr. Sterling Hilton, and Dr. Ross Larsen. You have helped me in so many ways that I can't count them all. From agreeing to be a part of the committee to the guidance, suggestions, and support you've provided along the way, your contributions are appreciated and will never be forgotten.

I gratefully acknowledge the Academic Vice President's office that made this work possible. It was a great learning experience that will be useful for better decision making. I learned far beyond statistics and psychometrics; I learned real life applications.

My time at BYU was made enjoyable in large part for the many friends and groups that became part of my life. I am grateful for the EIME and IP&T students who made this experience enriched with friendship and fruitful discussions.

I am grateful for my family, especially my mother, for their support, prayers, and endless love helped me survive graduate school in the best and toughest times. I am grateful for my daughters Taj and Masa, for their unremitting inspiration and unconditional love that kept me awake and working through the long nights.

Finally and most importantly, I would like to thank and dedicate this work to my wife, Rasha. There are no words in this world that can express the love, support, and devotion she has

shown to me throughout the years. Her unwavering love, patience, and encouragement have been the foundation in which we have built our lives on these past five years. Her sacrifices and tolerance of me through these agonizing years has been my inspiration and motivation to move forward with my education and to further my career. Understanding me best as a Ph.D. herself has allowed me to succeed and it is to her I say thank you for everything.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: Introduction

Student evaluations of teaching effectiveness (SET) are commonly used to evaluate teaching and instructional quality. Ting (2000) argues that course ratings are expressions of students' perceptions of the educational environment, and such perceptions are the outcome of an interaction process between students and teachers in the classroom. SET are used to inform decisions about teacher effectiveness in most colleges and universities. The results are used to provide feedback to faculty in hopes fostering teacher improvement. They are also used to monitor quality, to inform promotion and tenure decisions. In addition, the results are often used to provide information to students as they select teachers and courses (Boysen, 2015a; Marsh, 2007; Marsh, Ginns, Morin, & Nagengast, 2011; Rantanen, 2013).

Numerous publications have been written about the validity and reliability of SET ratings and the effect of different factors that influence such ratings (Boysen, Kelly, Raesly, & Casner, 2014; Dodeen, 2013; Marsh et al., 2011; Narayanan, Sawaya, & Johnson, 2014; Rantanen, 2013). Furthermore, studies have shown that SET ratings can be reliable and are a function of the teacher rather than other class variables. Variables related to potential sources of bias in the ratings such as class grades or class size tend to have a relatively minor effect on an instructor's overall SET rating (DeFrain, 2016; Dodeen, 2013).

However, less research has been devoted to the use and interpretation of SET ratings. Boysen (2015b, p. 151) argues that in higher education, SET are the source of an "eternal debate" even with vast research regarding their validity. For example, composite means of student ratings at the class-level are commonly interpreted and reported without understanding their meaning. Franklin (2001) concludes that it is a common mistake to consider raw mean scores a precise measure of SET. Boysen (2015a) states that "means are only an estimate of true

scores; thus, teaching evaluation means should be interpreted as estimates falling within a possible range of scores rather than a representation of true teaching competency" (p.151). SET in higher education typically have a negatively skewed distribution because students are likely to select higher ratings more often than lower ratings (Nulty, 2008; Zumrawi, Bates, & Schroeder, 2014). Thus, a margin of error based on central limit theorem (CLT) is inappropriate and undesirable when ratings have skewed distributions. Determining the appropriate unit of analysis is another concern in SET research. The vast majority of SET research studies use the class as the proper unit of analysis (Lüdtke, Robitzsch, Trautwein, & Kunter, 2009; Marsh, 2007; Morin, Marsh, Nagengast, & Scalas, 2014). Moreover, it has been shown that how it is valuable to incorporate a multilevel perspective into the use and interpretation of SET results (Lang & Kersting, 2007).

Need For the Study

This research fills a gap in the SET literature by evaluating (a) reliability for aggregated student ratings, and (b) alternative methods of describing the degree of uncertainty associated with estimates of the mean composite rating of instructor effectiveness averaged across the number of responding students in each class. To date, there have been few, if any, studies to recommend a method which provides accurate results from the psychometric properties of the aggregated ratings on class level. Furthermore, very few studies touch on the topic of reporting reliability for aggregated student ratings (Lang & Kersting, 2007; Lüdtke et. al., 2009). Additionally, the bulk of the literature focuses primarily on validity, reliability, multilevel analysis, and descriptive studies of student ratings (Beran & Violato, 2005; Campbell, 2005; Fernandez, Mateo, & Muniz, 1998; Hobson & Talbot, 2001; Kulik, 2001; Liu, 2012; Ory, 2001; Ory & Ryan, 2001; Pincus & Schmelkin, 2003; Radmacher & Martin, 2001; Sojka, Gupta, &

Deeter-Schmelz, 2002). Less research has been devoted to the use and interpretation of SET ratings to provide accurate results by developing inferential procedures about the mean rating (Franklin, 2001; James, Schraw, & Kuch, 2015; Miller & Penfield, 2005; Penfield & Miller, 2004).

We evaluated four methods of describing the degree of uncertainty associated with estimates of the mean composite rating of instructor effectiveness averaged across the number of responding students in each class. The four methods are (a) a logit transformation; (b) a resampling; (c) a non-parametric Bootstrapping (accelerated bias-corrected BCa); and (d) Bayesian bootstrap.

Purpose

The purpose of this study was threefold: First, to examine how the reliability of mean SET ratings varies as a function of the number of responding students and the number of items. Second, to evaluate four alternative methods of describing the degree of uncertainty associated with estimates of the mean composite rating of instructor effectiveness averaged across the number of responding students in each class. Third, to examine how the degree of uncertainty in class means varies as a function of the number of responding students.

Research Questions

1. What is the reliability of the estimated class means for the composite instructor effectiveness ratings?
    a. How does the estimated reliability vary as a function of the number of items and whether the items are classified as fixed or random?
    b. How does the estimated reliability vary as a function of both the number of items and the number of respondents?

   c. What minimum number of respondents should be required before reporting the class mean from single semester/term as an indicator of an instructor's effectiveness?

2. How do the various methods of estimating the range of uncertainty in class means compare in terms of the following concerns?
   a. The width of the interval.
   b. The degree and direction of asymmetry of the interval about the mean.
   c. The proportion of interval replications which contain the universe mean.
   d. The feasibility of the estimation procedures (i.e., how likely is it that algorithmic procedures for estimating the means of all classes across the university can be automated and systematized into a set of procedures that can be routinely implemented each semester or term with minimal human oversight and available computer resources?
3. How do the different methods of estimating the degree of uncertainty vary as a function of the number of responding students?

Definition of Terms

Observed class mean. The mean of the teacher ratings averaged across all of the responding students in a particular class. The observed class mean is a random variable. It is expected to vary somewhat from one sample of students to another and from one measurement occasion to another. Therefore, the reported mean obtained from any one sample of students on a single rating occasion is an estimate of the universe mean that would result if we could obtain ratings from all students in the class on all possible rating occasions. However, the observed class mean provides a basis for making reasoned inference about the unknown universe mean.

Generalizability theory (Cronbach, Gleser, Nanda, & Rajarantnam, 1972) provides a basis for investigating the accuracy of such inferences. In the SET context, students typically are only asked to rate their class instructor once during a given term or semester, but the ratings obtained from the students who responded on that single occasion are presumed to be representative of the ratings that would have been obtained if the ratings had been collected on any other acceptable rating occasion during that term.

Universe score. The universe score for a person is defined as the mean score for that person averaged over all included conditions of each facet in the *universe of admissible observations* (Figure 1A, 1B, and 1C). Universe scores are analogous to true scores in Classical Test Theory and have a similar meaning. In G-theory they are not called true scores because each examinee is expected to have a different universe score for each universe of admissible observations. The universe score for a given person is expected to vary from one universe of admissible observations to another depending upon which facets are included and the number of conditions within each facet as shown in Figure 1D.

Figure 1C shows the simplest design for which G-theory can be used. G-theory requires one factor that functions as the object of measurement plus at least one factor that is a facet. The reliability-like g-coefficient obtained by applying G-theory to data from this design will be equal to Cronbach's alpha coefficient and Hoyt's reliability coefficient from Classical Test Theory. This equality is not true for any other designs.

The last design (Figure 1E) is often used in informal rating situations, but it is an extremely weak design that should be avoided in conducting research and any time the goal is to obtain generalizable ratings. Raters are not a facet of this design because there is only one rater per person. Similarly, Task is not a facet because there is only one task per person. The Rater and

A. Three-facet Design (Person x Task x Occasion by Rater)

	Task 1								Task 2								Estimated
	Occasion 1				Occasion 2				Occasion 1				Occasion 2				Universe
person	R1	R2	R3	R8	R1	R5	R6	R8	R1	R5	R6	R8	R1	R5	R6	R8	Score
1	8	9	8	8	5	8	6	8	8	7	8	8	8	7	9	8	7.6875
2	7	5	6	5	6	7	6	4	8	8	5	6	7	7	5	6	6.1250
3	9	9	7	7	8	9	7	7	7	9	8	8	8	7	7	8	7.8125
4	7	7	7	4	4	8	4	4	4	9	7	8	8	9	8	8	6.6250
5	8	6	9	7	6	7	7	7	5	7	7	8	6	9	7	9	7.1875
6	5	3	8	1	4	3	6	1	7	9	9	6	7	7	6	6	5.5000
7	5	3	3	3	4	4	4	3	5	5	4	4	4	6	5	4	4.1250
8	2	2	4	2	2	2	2	2	4	5	4	2	2	5	4	2	2.8750
9	5	3	5	4	3	4	4	4	6	6	6	6	5	4	6	6	4.7500
10	8	7	6	2	7	5	5	2	6	6	4	3	5	6	4	3	5.0625
11	4	5	6	1	6	3	3	2	6	6	4	4	6	4	4	4	4.3750
12	8	7	6	6	6	6	6	6	7	9	6	6	7	9	7	7	6.8125

B. Two-facet Design (Person x Task x by Rater)

	Occasion 1								Estimated
	Task 1				Task 2				Universe
person	R1	R2	R3	R8	R1	R5	R6	R8	Score
1	8	9	8	8	8	7	8	8	8.000
2	7	5	6	5	8	8	5	6	6.250
3	9	9	7	7	7	9	8	8	8.000
4	7	7	7	4	4	9	7	8	6.625
5	8	6	9	7	5	7	7	8	7.125
6	5	3	8	1	7	9	9	6	6.000
7	5	3	3	3	5	5	4	4	4.000
8	2	2	4	2	4	5	4	2	3.125
9	5	3	5	4	6	6	6	6	5.125
10	8	7	6	2	6	6	4	3	5.250
11	4	5	6	1	6	6	4	4	4.500
12	8	7	6	6	7	9	6	6	6.875

C. One-facet Design (Person by Rater)

	Task 1				Estimated
					Universe
person	R1	R2	R3	R8	Score
1	8	9	8	8	8.25
2	7	5	6	5	5.75
3	9	9	7	7	8.00
4	7	7	7	4	6.25
5	8	6	9	7	7.50
6	5	3	8	1	4.25
7	5	3	3	3	3.50
8	2	2	4	2	2.50
9	5	3	5	4	4.25
10	8	7	6	2	5.75
11	4	5	6	1	4.00
12	8	7	6	6	6.75

D. Summary of Universe Scores

	Summary of Estimated Universe Scores		
Person	PxTxOxR	PxTxR	PxR
1	7.6875	8.000	8.25
2	6.1250	6.250	5.75
3	7.8125	8.000	8.00
4	6.6250	6.625	6.25
5	7.1875	7.125	7.50
6	5.5000	6.000	4.25
7	4.1250	4.000	3.50
8	2.8750	3.125	2.50
9	4.7500	5.125	4.25
10	5.0625	5.250	5.75
11	4.3750	4.500	4.00
12	6.8125	6.875	6.75

E. No Facet Design

	T1
person	R1
1	8
2	7
3	9
4	7
5	8
6	5
7	5
8	2
9	5
10	8
11	4
12	8

Figure 1. Alternative universes of admissible observations and estimated universe scores.

Task variables are completely confounded, and the effects of each cannot be disentangled or estimated separately. It is impossible to estimate the dependability of the ratings obtained from this design.

The one-, two-, and three-facet designs can each be expanded to include more persons (examinees), more tasks, more raters, or more occasions. Increasing the number of conditions in any of these facets will not change the structure of the design, but it will change the universe of admissible observations. Consequently, such changes will also likely change the estimated universe scores and the reliability of those estimates.

Universe class mean. The mean that would result if the University could obtain ratings from all students in a particular class on all possible rating occasions in a given semester.

Variance. Variance is a measure of dispersion. In general, the variance of a group of scores or observations is a single number which summarizes the degree to which the scores within that group are dispersed about (i.e., spread out away from) their mean. Unless the scores in a group are all the same, some scores will be less than the mean, and other scores will be greater than the mean. The variance is influenced by the location of every score relative to the group mean.

The variance of a set of scores is computed by calculating the squared deviation of each observed score from the group mean and then calculating the average of those squared deviations. The resulting number can be interpreted as a measure of the inconsistency or variability among the scores in the group. If the scores within a group are all the same, then the variance will be zero, and the scores will be perfectly consistent. The more the scores are spread out away from the group mean, the larger the variance will be.

In the SET context, it is possible to compute the variance at multiple levels of aggregation in the data hierarchy.

1. At the lowest level in the hierarchy (i.e., the item level), a variance can be computed which summarizes the inconsistency among the responses of a single student to the various items in the SET questionnaire.
2. At the next highest level in the hierarchy (i.e., the student level) a variance statistic can be computed summarizes the variability of the mean rating obtained from each student about the mean of all the students in the class. When the mean rating of the different students are clustered closely about the class mean, this variance will be small. The farther the

student means are spread out away from the class mean, the greater the inconsistency among the students' ratings and the larger the variance will be.

3. A third variance statistic can be computed at the department level. The variance at this level describes the variability or inconsistency among the means of the different classes in that department during that semester.

Variance within-group. Software programs used to conduct multilevel modeling do not directly compute the variance of the student means within each class about their class mean. Instead, the software computes and reports a single, global estimate of the variance in the different classes. This pooled estimate is a weighted average of all the individual class variances. Consistent with the published literature on multilevel modeling, in this study this estimated variance component is labeled the within-group variance and is symbolized by σ_{WG}^2. This variance describes the average variability of the student means within any given class about the mean of that class.

Variance between-groups. Multilevel software programs also do not directly compute the variability of all the classes within a department about the mean of that department. Instead, these programs compute a single, global estimate of the average variance of the degree to which the means of the different classes within a department are dispersed about the mean of that department during in a particular semester. This estimated variance component is labeled the between-group variance and is symbolized by σ_{BG}^2.

The standard error of the mean. The standard deviation of the sampling distribution of the observed class mean.

Confidence interval. A confidence interval (CI) defines a range within which the unknown universe mean for a given class can be said to lie, with a given level of confidence.

Reporting confidence intervals along with the estimated class means emphasizes that the reported mean is a fallible estimate of the unknown true mean for a given class and provides a description of how accurate (or inaccurate) that estimate is.

Confidence level. A confidence level specifies how confident we are in our estimate of class means. If a 90% confidence level is selected, 90 out of 100 samples will have the unknown true population mean within the range of precision specified.

Credible interval. The Bayesian analog of a confidence interval is a credible interval. The credible interval is calculated from the posterior distribution to quantify uncertainty about the unknown true class mean. A 95% credible interval is one that has a 95% chance of containing the unknown true class mean.

Intraclass correlation coefficient. An interclass correlation coefficient is an inferential statistic that can help to determine whether or not a hierarchical model is necessary. It can also help to understand how much of the overall variation in the response is explained simply by clustering.

CHAPTER 2: Literature Review

Introduction

Student evaluations of teaching effectiveness (SET) are essential in any higher educational institution. SET are used to inform decisions about teacher effectiveness in most colleges and universities. The results are used to provide feedback to faculty in hopes fostering teacher improvement. They are also important because they offer feedback that impacts the instructor's self-image and professional satisfaction.

Numerous publications have been written about the validity and reliability of SET ratings and the effect of different factors that influence such ratings. However, less research has been devoted to the use and interpretation of SET ratings. What is the reliability of aggregated students ratings? Which method is appropriate to describe the degree of uncertainty associated with estimates of the mean composite rating of instructor effectiveness averaged across the number of responding students in each class?

In an attempt to answer these questions, three different areas of research will be reviewed: (a) multilevel nature of SET and unit of analysis, (b) psychometric properties of aggregated student ratings, and (c) estimation of class mean and range of uncertainty.

Multilevel Nature of SET and Unit of Analysis

Many scholars have asserted that SET are a multilevel phenomenon in which students are nested within classes determining the proper unit of analysis based on this perspective a critical methodological concern (Lang & Kersting, 2007; Lüdtke et al., 2009; Lüdtke, Trautwein, Kunter, & Baumert, 2006; Marsh, 2007; Marsh, Lüdtke, Nagengast, Trautwein, Morin, Abduljabbar, & Koller, 2012; Raudenbush & Bryk, 2002; Schweig, 2013; Schweig, 2016; and Ting, 2000). Marsh et al. (2012) argued that the multilevel structure of educational data is

appropriately represented by “multilevel modeling and related techniques” (p. 108). Morin et al. (2014) argued that the multilevel structure should not rely on manifest-variable models but should consider latent-variable models “that control measurement error at L1 and L2, and sampling error in the aggregation of L1 ratings to form L2 constructs” (p. 1). Other studies argued that in educational research both Level 1 (individual-level) and Level 2 (class-level) have been considered as the unit of analysis and this decision depends on the research question (Lüdtke et al., 2009; Marsh et. al, 2012). However, in classroom research, it is important to distinguish between Level 1 which is based on student responses and Level 2 (e.g., teacher or class) which is based on aggregation of student responses within the class (Marsh et al., 2012; Morin et al., 2014).

A review of the literature by Marsh (2007) reported many studies recommend using class-level as the unit of analysis. Marsh (2007, p. 329) concluded that there is a clear consensus in SET research that the class-average or individual teacher mean is the appropriate unit of analysis, rather than the individual [student] level. Lang and Kersting (2007) stated that “it is particularly important to determine the appropriate level of analysis, as an aggregated variable at a higher-order level might well measure a different construct than does its namesake at the individual level” (p. 193). Moreover, Lüdtke et al. (2006) argued that the aggregation of student ratings at class-level is the recommended approach in SET studies because the aggregation reflects the “perceptions of the shared learning environment” (p. 216).

Morin et al. (2014) discussed the consequences of ignoring the hierarchal structure of classroom data and failing to use the class-level as unit of analysis. They argued, “failure to analyse the data at the proper level of analysis would lead the researcher to conclude that the effects are located at the individual level when in fact they are located at the classroom level” (p.

6). Consequently, a systematic bias may result when ignoring the level-2 because the effect of students level-1 residual.

Psychometric Properties of Aggregated Student Ratings

Many scholars have emphasized the importance of investigating the psychometric properties of the aggregated ratings among students in a class (Gross, Lakey, Edinger, Orehek, & Heffron, 2009; Lüdtke et al., 2009; Lüdtke et al, 2006; Marsh & Roche,1997; Marsh et al., 2012; Praetorius, Lenske, & Helmke, 2012; Schweig, 2016). Lüdtke et al. (2009) argued that "before aggregating student perceptions of learning characteristics at the class or school level, researchers must determine whether it makes sense to form an aggregate variable" (p. 122). In their review of the literature, Lüdtke et al. (2006) concluded that to justify and assess aggregated student ratings at the class level, the researcher should evaluate the reliability of aggregated student ratings. Nelson and Christ (2016) argued that "class-level reliability and agreement indices supplement more traditional considerations, such as test–retest and internal reliability of scales. The interpretation of the aggregate, or mean, of class ratings, requires evidence that the aggregate is a reliable indicator of consensus" (p. 421).

Evaluating the reliability of aggregated student ratings. Two different ways to determine whether aggregated student ratings are reliable indicators of the group-level construct are considered in the literature review: (a) generalizability coefficients, and (b) intraclass correlation coefficients.

Generalizability theory. Generalizability theory (Cronbach et al., 1972) is a measurement theory in which the error variance in a set of ratings or observations is partitioned into two or more components representing different likely sources of measurement error. Generalizability theory extends classical test theory in a way that permits researchers to generalize about a

person's observed behavior (Proctor, 2003). Cronbach et al. (1972) stated a key idea in generalizability theory as:

> A behavioral measurement is a sample from the collection of measurements that might have been made, and interest attached to the obtained score only because it is [presumed to be] representative of the whole collection or universe. If the decision-maker could, he would measure the person exhaustively and take the average over all the measurements. (p. 18)

The sources of error variation are a set of similar conditions within the acceptable universe that are called facets (Brennan, 2001). A *facet* is a potential source of measurement error or inconsistency in the ratings. Researchers decide which facets should be included in the universe to which they intend to generalize (Brennan, 2001). In the context of SET, both students and items are considered important facets (Ibrahim, 2011). Rating occasion is another facet that could be included but usually is not. The universe score "replaces the true score within classical test theory and places emphasis on the idea that there are many universes to which a researcher can generalize" (Proctor, 2003, p. 13). The term *universe* is used because there are different universes to which the researcher may want to generalize (Brennan, 1992). According to (Briesch, Swaminathan, Welsh, & Chafouleas, 2014) generalizability theory "offers increased utility for assessment research given the ability to concurrently examine multiple sources of variance, inform both relative and absolute decision making, and determine both the consistency and generalizability of results" (p. 13).

Generalizability theory distinguishes between two types of error variance: (a) relative error variance and (b) absolute error variance. The relative error variance is applicable when a researcher wants to make decisions about the relative standing of persons to each other and

whether each individual is above or below the group mean (Briesch et al., 2014). Relative error is the difference between a particular person's observed deviation score and his/her universe deviation score (Brennan, 2001). In contrast, absolute error is the difference between a person's observed score and his/her universe score (Brennan, 2001). The absolute error variance is used when the researcher wants to make inferences about an individual's standing comparted to some predetermined level of acceptable performance such as a cutscore that is intended to define mastery.

Generalizability theory includes two types of studies: (a) a generalizability or G-study, and (b) a decision or D-study. The G-study is used to obtain estimates of variance components for the object of measurement and for each facet and interaction in the model (Brennan, 2001). The estimated G-study variance components are then analyzed in a subsequent D-study. The purpose of the D-study is to compute generalizability coefficients and to project how those coefficients are likely to vary if the number of levels of one or more facets (e.g., items or raters) is increased or decreased. Ibrahim (2011) argued that the D-study is important in making decisions about the number of items and students needed to obtain "dependable ratings" (p. 254). Alkharusi (2012) argued that the D-study addressed the question "What should be done differently if you are going to rely on this measurement procedure for making future decisions?" (p. 193)

Two different kinds of generalizability coefficients can be calculated within the generalizability theory framework: (a) a generalizability coefficient in which the relative standing of objects of measurement are of concern and (b) an index of dependability in which the absolute standing of the objects of measurement are of concern. The generalizability coefficient (Brennan, 1992) is defined as:

$$E\rho_{\sigma}^{2} = \frac{\sigma_{\tau}^{2}}{\sigma_{\tau}^{2} + \sigma_{\delta}^{2}} \quad (2.1)$$

In (2.1) σ_{τ}^{2} is the variance of the object of measurement and σ_{δ}^{2} is the relative error variance. This generalizability coefficient in "a conceptual sense are related to the traditionally used CTT reliability coefficient" (Raykov & Marcoulides, 2006, p. 84).

In SET literature, reliability estimates for aggregated measures such as the class level are "relatively neglected" (Wei & Haertel, 2011, p. 15). Assessing the reliability of class means is needed to examine how well interactions can be distinguished based on ratings given by individual student ratings (Schweig, 2013). Unfortunately, many previous studies have estimated Cronbach's alpha reliability and ignored the psychometric quality of their aggregated constructs (Jeon, Lee, Hwang, & Kang, 2009; Lüdtke et al., 2009). The consequences of ignoring the reliability of aggregated measures may lead to the misinterpretation or misuse of scores (Jeon et. al, 2009). Accordingly, it is important before using aggregated scores at the class level to investigate the psychometric properties of those scores. This examination is very important to "making accurate, substantial inferences based on those scores" (Jeon et al., 2009, p. 149).

Regarding the reliability of class means, Kane and Brennan (1977) used a split-plot design to examine the reliability of class means with students nested in classes and items crossed with students. Four different reliability coefficients examined based on different conditions: (a) infinite universe of students and items, (b) infinite universe of students and fixed set of items, (c) infinite universe of items and fixed set of students, and (d) fixed set of students and fixed set of items. These four generalizability coefficients have been compared with three coefficients for estimating the reliability of class means (a) Wiley's coefficient which is equivalent to infinite universe of students and fixed set of items; (b) Thrash and Porter's coefficient which is equivalent to infinite universe of items and fixed set of students; and (c) Shaycoft's coefficient

which is equivalent to an upper bound for infinite universe of items and fixed set of students, and lower bound for fixed set of students and fixed set of items. Kane and Brennan (1977) concluded that the most appropriate reliability coefficient is an infinite universe of students and item. Other studies used the variations of split plot design at the class level by using different conditions for items and students (e.g., Ibrahim, 2011; Kane, Gillmore, & Crooks, 1976; Wei & Haertel, 2011). However, the decision to treat facets as random or fixed can only be made in the context of the study (Brennan, 2011; Schweig, 2013). Fixing a facet in the design leads to large generalizability coefficient because it "restricts the universe of generalization and, in doing so, decreases the gap between observed and universe scores at the price of narrowing interpretations" (Brennan, 2011, p. 12).

A number of studies have used multilevel models to estimate the different variance components to be used in the generalizability theory instead of using ANOVA (e.g., Geldhof, Preacher, & Zyphur, 2014; Jeon et al., 2009; Raykov & Marcoulides, 2006). In their study, Jeon et al. (2009) investigated different methods of estimating the reliability of school-level scores using generalizability theory and multilevel models. They found that both methods provide very similar reliability estimates of school-level scores. Moreover, they argued that the multilevel models "offer many advantages in examining the relationships among individual-level and school-level measures" (p. 150). Raykov and Marcoulides (2006) used structural equation model for purposes of estimating the relative generalizability coefficient in one-faceted and two-faceted crossed designs. Their results showed the aspects of commonality of the structural equation modeling and the generalizability theory. Geldhof et al. (2014) discussed the relationship between generalizability theory facets and parallel elements of the multilevel confirmatory factor analysis model and concluded that the multilevel confirmatory factor analysis model provides

decomposition of a scale variance similar to a parallel model derived from generalizability theory.

Intraclass correlation coefficients. Many studies have used intraclass correlation ICC_1 and ICC_2 to determine whether aggregated individual-level ratings are reliable indicators of group-level constructs (Frenzel, Goetz, Lüdtke, Pekrun, & Sutton, 2009; LeBreton & Senter, 2008; Lüdtke et al., 2009; Marsh et al., 2012; Newman & Sin, 2007). Woehr, Loignon, Schmidt, Loughry, and Ohland (2015) stated that ICC_1 describes the amount of variance in each item that can be attributed to belonging to the class-level. Moreover, LeBreton and Senter (2008) argued that "within the context of multilevel modeling, the ICC_1 is typically used to provide an estimate of effect size indicating the extent to which individual ratings (e.g., climate ratings) are attributable to group membership" (pp. 833-834). In contrast, ICC_2 provides a reliable estimate of the class-level group means. ICC_1 formula is defined as:

$$ICC_1 = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2} \tag{2.2}$$

In (2.2) σ_B^2 is the between-group variance for the observed item, and σ_W^2 is the within group variance for the observed item, and *k* is the class size. ICC_2 formula is defined as:

$$ICC_2 = \frac{k \sum ICC_1/m}{1 + (k-1)\sum ICC_1/m} \tag{2.3}$$

In (2.3) *k* is the average class size and *m* is the number of items.

Lüdtke et al. (2009, p. 123) argued that as the number of students increases, the reliability of the class-mean rating as estimated by the ICC_2 increases. In other words, the more students in class who provide ratings, the more accurately the class-mean rating will reflect the true value of the construct being measured. The interpretation of ICC_1 and ICC_2 indexes are common within

the research literature. However, the rule of thumb for interpreting is “often implemented with minimal scrutiny and are sometimes mischaracterized” (Woehr et. al, 2015, p. 707).

In their literature review, Woehr et. al (2015) discussed different rules of thumb which have applied in previous research to interpret ICC_1 . For example, some researchers used rule of thumb (median = 0.12, range = 0.00-0.50). Other researchers referenced Bliese (2000) rule of thumb (0.05 to 0.20, never exceeding 0.30). On the other hand, LeBreton and Senter (2008) suggested another approach to interpreting values for ICC_1 by adopting the effect size way of interpretation. For example, a value of .01 might be considered a small effect, a value of .10 might be considered a medium effect, and a value of .25 might be considered a large effect.

LeBreton and Senter (2008) declared that different research questions are answered by ICC_1 and ICC_2. ICC_1 “informs a researcher as to whether judges’ ratings are affected by group membership, thus, a question should drive the use of ICC_1 “(e.g., Does group membership affect judges’ ratings?)” (p. 834). On another hand, ICC_2 tells a researcher how reliably the mean rating distinguishes between groups, thus, a question should drive the use of “(e.g., do judges’ mean ratings reliably distinguish among the groups/targets?)” (p. 834).

Uncertainty in Estimates of the Universe Mean for a Class

To date, there have been few, if any, studies which have attempted to describe the degree of uncertainty in the estimates of class mean. In fact, very few studies have been devoted to the use and interpretation of the SET results (Franklin, 2001; James et al., 2015; Miller & Penfield, 2005; Penfield & Miller, 2004). On the other hand, the bulk of the literature focuses primarily on validity, reliability, multilevel analysis, and descriptive studies of student ratings (Beran & Violato, 2005; Campbell, 2005; Fernandez, Mateo, & Muniz, 1998; Hobson & Talbot, 2001;

Kulik, 2001; Liu, 2012; Ory, 2001; Ory & Ryan, 2001; Pincus & Schmelkin, 2003; Radmacher & Martin, 2001; Sojka et al., 2002).

Boysen (2015a, p. 151) argued that in higher education, SET are the source of "eternal debate" even with vast research in their validity. For example, using class-level as an appropriate unit of analysis, composite means of student ratings at this level are commonly interpreted and reported without understanding their meaning. Furthermore, Boysen (2015a) stated that "means are only an estimate of true scores; thus, teaching evaluation means should be interpreted as estimates falling within possible range of scores rather than a representation of true teaching competency" (p.151). Boysen (2016) warned against the tendency of SET users to overinterpret the observed class mean. He declared that the "cardinal rule" (p. 279) should be to report a confidence interval along with the observed mean from each class in order to emphasize that the observed mean is an estimate of the unknown universe mean. Abrami (2001), Franklin (2001), and Boysen et al. (2014) provide further support for this recommendation.

Confidence Interval with Skewed SET Distribution

The confidence interval (CI) of a sample mean is an important sample statistic that is intended to help users qualify the inferences which they make about the population from which this sample was drawn. The sample mean provides a point estimate of an unknown population parameter. However, an interval estimate "specifies a range of values on either side of the sample statistic within which the population parameter can be expected to fall with a chosen level of confidence" (Sim & Reid, 1999, p. 189). In the context of this study, a confidence interval is an estimated range of values within which the true mean of a class has a 95% probability of being located. The confidence interval for each class is based on the mean and the variability of the ratings obtained from the students in that class and the number of students who

responded. The resulting confidence interval for each class is a random variable. The location and the width of the interval will vary if a different sample of ratings is obtained.

Confidence intervals can be very useful, but they are frequently misinterpreted. For example, it is incorrect to compute a 95% confidence interval and then declare that one can be 95% certain that the true mean lies within that interval. After a confidence interval has been constructed, it either does or does not encompass the true mean. Hence, the probability that is captures the mean is either 0 or 1.0 (Cumming & Fidler, 2009; Sim & Reid, 1999).

In order to correctly interpret a confidence interval, it is helpful to consider not only the interval obtained from the observed sample, but other intervals which could have been constructed based on information from other samples of ratings that could have been collected. For instance, Figure 2 displays examples of 50 different confidence intervals each based on a different sample. The vertical line in this diagram represents the universe mean. It is important to note that the universe mean is a constant. It is not a random variable which varies from sample to sample. In reality, the value of this parameter would not be known, but a fictitious example of such a parameter is reported here for pedagogical purposes. The location and width of each interval in Figure 2 differs depending on the information in the corresponding sample of ratings. All of these intervals are based on samples of the same size. The location of each interval is determined by the mean of corresponding sample, and the width of each interval is influenced by the variability of each ratings in the sample. If the ratings in any one of the 50 samples had been different, then the confidence interval for that sample would be different.

Not that 48 (96%) of the 50 intervals capture the true mean, and 2 (4%) of them do not. Freedman, Pisani, Purves, and Adhikari (1991) use an interesting analogy to describe the process of estimating the unknown population mean based on an accompanying confidence interval.

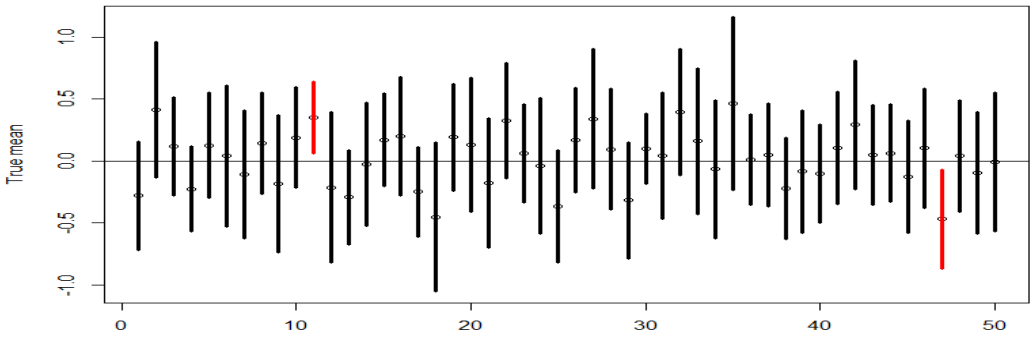

Figure 2. The population mean and 95% confidence interval for 50 samples.

They point out that if a researcher takes repeated samples and computes the sample mean and a confidence interval for each sample, some of the resulting intervals will capture the universe value and some will not. They conclude that this process is "like buying a used car. Sometimes you get a lemon—a confidence interval which doesn't cover the parameter" (Freedman et al., 1991, p. 351). This conclusion should serve as a warning to professors and administrators who interpret confidence intervals in the context of SET. Not all reported intervals will include an instructor's universe mean. All that can be said with confidence is that in the long run the procedure will work. If we take many repeated samples and compute a confidence interval for each sample, in the long run 95% of the resulting intervals will encompass the parameter that is being estimated.

If the proper steps for computing the sample mean and confidence interval are followed for each sample, then what can be said with a 95% level of confidence is that the procedure by which the intervals were produced will capture the unknown true mean in 95% of cases. It is the method or process that the user can be confident about.

Several methods have been described in the literature for estimating the confidence interval of the population mean. Most widely used methods designed to estimate confidence interval based on large sample sizes and assuming that the observations are normally distributed, which most SET data do not follow. It is known that confidence intervals for the population mean of a skewed distribution with small sample sizes usually have poor coverage rate (Meeden, 1999; Shi & Golam Kibria, 2007). Therefore, numerous studies in the literature have presented other methods (e.g., a transformation method, bootstrap method) to obtain acceptable coverage rate and small interval width with skewed distribution and small sample sizes (Agresti & Coull, 1998; Calzada & Gardner, 2011; Ghosh & Polansky, 2016; Liu, 2009; Meeden, 1999; Newcombe, 1998; Shi & Golam Kibria, 2007; Willink, 2007; Zhou & Gao, 1997; Zou & Donner, 2008).

Other studies have highlighted another issue associated with estimating the confidence interval for a bounded parameter. Chen (2008) argued that intolerable error could emerge when constructing a confidence interval for bounded variables from skewed distribution even with large sample size. Sappakitkamjorn and Niwitpong (2013) argued that such error,

> is due to the fact that the information regarding the restriction is simply ignored. It is, therefore, of significant interest to construct confidence intervals for the parameters that include the additional information on parameter values being bounded to enhance the accuracy of the interval estimation. (p. 1416)

SET in higher education often have a negatively skewed distribution because students are prone to select higher ratings more often than lower ratings (Nulty, 2008; Zumrawi, Bates, & Schroeder, 2014). Estimating a margin of error using methods that are based on the central limit theorem (CLT) is inappropriate and suffer from several deficiencies including the small number

of raters, the bounded nature of the rating scale items, and the standard error represented in the measurement (Boysen, 2016; Miller & Penfield, 2005). James et al. (2015) discussed two types of error related to a small number of raters including:

1. Error caused by sample-to-sample variation, this error related to the standard deviation of sample mean. Lower precision about the SET score is expected when the standard deviation is large
2. Error caused by sampling bias, this error related to the selection bias that makes a difference between true population mean and observed mean.

Therefore, it is important to construct a confidence interval of a sample mean that is not limited by the assumption of population normality and less affected by the bounded nature of rating scale items (Miller & Penfield, 2005). In addition, researchers should provide information that show how much confidence one should have that the unknown true population mean lies within a certain distance from the obtained sample mean (Penfield & Miller, 2004).

Few published SET studies have paid attention to developing inferential procedures that can provide more information about the lack of precision associated with using the observed class mean as an estimate of the universe mean. Penfield and Miller (2004) proposed an asymmetric score confidence interval for the population mean of a rating scale variable which can be used with small sample sizes. Franklin (2001) suggested an approach of estimating the margin of error associated with item's mean. However, the two methods have some limitations: (a) the confidence interval is obtained for the item's mean but it does not apply for the aggregated item's means (mean of the means) of a class; and (b) the methods assumes the rating has a symmetrical distribution.

James et al. (2015) proposed a method to determine the margin of error around a SET mean score as a function of class size, response rate, and sample variability. They used a statistical sampling theory explained by Govindarajulu (1999) and Barnett (2002) with some modification. However, this method has a number of limitations: (a) it assumes the distribution of sample means follows a Student's *t*-distribution, which is not the case with SET scores; and (b) it did not examine the bounded nature of rating scale items and the celling effect associated with very high means on a scale.

There have been few, if any, studies in the SET literature that recommend a method that provides accurate results from the psychometric properties of the aggregated ratings on class level. Method that address issues in SET such as a negatively skewed distribution, low response rate, and bounded distribution. Luh and Guo (2001) argued that "since violation of the normality assumption may be fairly common in applied research and since the consequences of non-normality for test statistics are difficult to investigate, robust and efficient alternatives to deal with the problem are needed" (pp. 227-228).

Logit transformation method. Several studies in the literature recommended the use of the logit transformation method to address violations of the normality assumption (Van Albada & Robinson, 2007; Williamson & Gaston, 1999). Other studies suggested that the logit transformation is appropriate when the data are bounded interval (Gart, Pettigrew, & Thomas, 1985; Hu, Yeilding, Davis, & Zhou, 2011; Lesaffre, Rizopoulos, & Tsonaka, 2007).

Choi, Fine, and Brookhart (2013) conducted a simulation to study the performance of a simple Wald-based CIs using transformation. Their findings indicated that with small to moderate sample sizes, the coverage of the Wald intervals has improved by transformation. Williamson and Gaston (1999) explained the process of using logit transformation as: (a) any

bounded data can be rescaled to a 0-1 scale and expressed as a proportion p; (b) q = 1 – p; (c) the standard form for the logit is log (p/q). However, Warton and Hui (2011) and Hu et al. (2011) argued that if the sample proportions equal to 0 and 1, this is problematic because it will transform to undefined values. Williamson and Gaston (1999) presented an ad hoc solution to this problem by adding "some small value ε to both the numerator and denominator of the logit function, which introduces minimal bias while still satisfying the criteria above" (p. 5). Hu et al. (2011) suggested another solution by transforming the data to "a slightly smaller interval of [δ, 1 - δ] with δ > 0 and then applying the logit transformation" (p. 499).

Bootstrap methods. Efron (1979) introduced bootstrap methods as a computer-based tool used to make inferences about unknown population parameters for which no assumptions are necessary regarding the underlying distribution. Bootstrap is a "procedure for estimating (approximating) the distribution of a statistics. It is based on resampling and simulation" (Mammen, 1992, p. 1). Efron and Tibshirani (1986, p. 54) argued that the bootstrapping methods are "a general methodology" to answer a question how accurate the sample mean as an estimator of the population mean. Bootstrap sampling is useful for quantifying the behavior of parameter estimates, such as its standard error, skewness, bias, or for calculating a confidence interval (Chihara & Hesterberg, 2011). Bootstrapping is an alternative method used when the distribution of the original sample is skewed and does not follow the central limit theorem (Ghosh & Polansky, 2016). Thus, several bootstrap methods were introduced in the literature to construct confidence intervals. However, Shi and Golam Kibria (2007) argued that "the accuracy of the bootstrap confidence interval depends on the number of bootstrap samples. If the number of bootstrap samples is large enough, the confidence interval may be very accurate for the specific sample" (p. 414). Briggs, Wonderling, and Mooney (1997) argued that the bootstrap method

"estimates the sampling distribution of a statistic through a large number of simulations, based on sampling with replacement from the original data. Confidence intervals can then be constructed using this empirical estimate of the sampling distribution" (p. 328).

Bootstrap methods are performed by following two approaches: (a) non-parametric; and (b) parametric. A non-parametric approach is more a general approach, which does not make an assumption about the form of distribution, but it can be computationally intensive (Carpenter & Bithell, 2000; Dixon, 2002; Kelley, 2005; Nicholls, 2014; O'Hagan, & Stevens, 2003). Moreover, Nixon, Wonderling, and Grieve (2010) argued that this approach involves re-sampling from the original sample with a replacement while preserving the original structure of the data (e.g., same size). On the other hand, the parametric bootstrap assumes that the random sample follows a specific distribution (Savoy, 1997).

Resampling. Resampling method is a percentile confidence interval for a population mean (Karian & Dudewicz, 2011). This non-parametric method "depends upon the fact that the empirical distribution function based on the bootstraps converges to the true distribution function asymptotically in sample size, and the empirical quantiles have a law of large numbers" (Puth, Neuhauser, & Ruxton, 2015, p. 893). Furthermore, this method does not have to make a distribution assumption, but it uses the distribution of the bootstrap sample statistic as a direct approximation of the data (Burn, 2003).

This method is widely used in the literature for many advantages. Carpenter and Bithell (2000) argued that "simplicity is the attraction of this method, and explains its continued popularity. Unlike the bootstrap-*t*, no estimates of the σ are required. Further, no invalid parameter values can be included in the interval" (p. 1152). On the other hand, Ghosh and Polansky (2014) argues that this method suffers from low coverage probability. To compute the

(1-α)100% confidence interval for the population mean, we take B resamples of size n with replacement from a population. Then, we use the (α/2)B^{th}, and the (1-α/2)B^{th} ordered bootstrapping means of all observed bootstrapping means as a lower and upper CI limits (Calzada & Gardner 2011).

Nicholls (2014) argued that this method performs well with bounded distributions and "neither do we have to worry about small sample size effects, or effective degrees of freedom. No mathematics required" (p. 910). Abu-Shawiesh, Banik, & Kibria (2011) used a simulation study to compare the percentile method to other bootstrap methods under different distributions and different sample sizes. They found that the percentile method performed better compared to other methods with non-normal distributions. However, other studies presented different results in which the percentile method perform worse compared to other methods (e.g., Ghosh and Polansky, 2014).

Ghosh and Polansky (2014) compare the bootstrap percentile CI to other methods (e.g., percentile *t*) through a simulation. They used three positively skewed distribution and three sample sizes (n = 15, 25, 50). The results showed that the bootstrap percentile performed poor with small sample size, but it provided more accurate coverage probability for large samples.

Non-parametric bootstrapping (Accelerated Bias-Corrected BCa). Efron (1987) introduced the BCa method to address the limitation of percentile method. Many researchers recommended using the accelerated bias-corrected percentile limit (BCa) method because it seeks to correct for skewness and bias in the bootstrap distribution (Calzada & Gardner, 2011; Carpenter & Bithell, 2000; Diciccio & Efron, 1996; Puth et. al., 2015). Calzada and Gardner (2011, p. 32) argued that the BCa procedure has "very desirable asymptotic characteristics" and its confidence interval "has stronger theoretical underpinnings and requires some easy to

program computation". In the non-parametric BCa two parameters are derived from the sample (a) the bias-correction constant z and (b) the acceleration a. The bias-correction constant z "adjusts the sampling distribution for the bias of the estimator," while the acceleration a "adjusts for the skewness of the sampling distribution" (Briggs et. al, 1997, p. 335). Diciccio and Efron (1996) argued that "the BCa method is an automatic algorithm for producing highly accurate confidence limits from a bootstrap distribution" (p. 192).Carpenter and Bithell (2000) argues that "this method generally has a smaller coverage error than the percentile" (p. 1154).

Tong, Chang, Jin, and Saminathan (2012) compared the behavior of four 95% bootstrap confidence intervals of different bootstrap methods including the bootstrap BCa under different distributions and sample sizes. They found that BCa with large sample sizes always results in higher coverage performance and shorter interval mean. Moreover, bootstrap methods including BCs perform better given non-normal dataset and small sample size. However, Wang (2001) argued that in non-parametric situations "an accurate estimator of *a* is not easy to obtain, and the BCa confidence sets may perform poorly" (p. 259).

In their simulation study, Banik and Kibria (2010) concluded that in case of small sample sizes and skewed distributions, the BCa method provides a narrow confidence interval width but it provides low coverage probability. However, their simulation considered only positively skewed distributions and did not consider bounded and negatively skewed distributions.

Bayesian bootstrap. The Bayesian bootstrap is a non-informative Dirichlet process introduced by Rubin (1981) as the Bayesian equivalent of the bootstrap. In the recent years, Bayesian inference has become widely studied in the literature (Alfaro, Zoller, & Lutzoni; 2003; Efron, 2015; Gu, Ghosal, & Roy, 2008; Huang, Li, Cheng, Cheung, & Wong, 2016; O'Hagan, & Stevens, 2003). Alfaro et al. (2003) explained the principle of the Bayesian as,

> In Bayesian inference one makes use of Bayes's theorem to condition inferences about the value of some parameter of interest on the observed data. Bayesian inference focuses on the quantity known as the posterior probability, defined as the probability of some hypothesis conditional on the observed data. The posterior probability is proportional to the product of the likelihood of the data, given that the hypothesis is correct and the prior probability of the hypothesis before any data have been collected. (p. 256)

The Bayesian bootstrap can be considered as doing the same like bootstrap except "the weights on the original data are resampled to give different posterior distributions" (Nicholls, 2014, p. 910). The weight may take any value between zero and *n* (sample size) and still sum to *n* (O'Hagan & Stevens, 2003). Moreover, the simulation is on the posterior distribution of the parameter rather than the sampling distribution of a statistic estimating a parameter (Rubin, 1981). In other words, the Bayesian bootstrap distribution is the posterior distribution of the parameter of interest (e.g., mean).

The interpretation of the Bayesian interval of the mean is different from the frequentist approaches. O'Hagan and Stevens (2003) argued that a "Bayesian 95% interval has the interpretation that there is a 95% probability that mean lies in the actual interval" (p. 39). They also stated that the Bayesian credible interval has "more direct and practically useful interpretation than a confidence interval" (p. 39).

O'Hagan and Stevens (2003) compared Bayesian bootstrap interval estimation for the mean with other methods (e.g., standard normal-theory based on Student *t* and non-parametric bootstrap). They used data comprise 26 observations that have lognormal distribution. Their results indicated that both non-parametric bootstrap and Bayesian bootstrap provided similar results and outperformed the Student *t* method. However, the Bayesian bootstrap provided

smoother distribution compared to the bootstrap distribution, and this would make the interpretation of the result more direct and informative. Similar conclusion was reported by other researchers (e.g., Lazar, Meeden, & Nelson, 2008; Lo, 1988; Weng, 1989).

Criteria for evaluating the different methods. The evaluation of the performance of confidence interval methods has been discussed widely in the literature (e.g., Abu-Shawiesh et al., 2011; Kelley & Rausch, 2006; Liu, 2009; Newcombe, 1998; Sappakitkamjorn & Niwitpong, 2013; Shi & Golam Kibria, 2007; Swift, 2009; Tong et al., 2012; Wu, Wong, & Jiang, 2003). Different criteria are considered as a criterion of the good estimators in choosing a confidence interval method, for example: (a) the coverage probabilities; (b) the average interval widths; (c) the coverage error; and (d) the upper and lower error probability.

The coverage probability represents the percentage of times that the actual parameter of interest (e.g., population mean) falls into the confidence intervals $[L, U]$ as $\Pr[L \leq \theta \leq U]$ where L and U are the lower and upper limits. Sim and Reid (1999) argued that the coverage probability of CI is indicative of its accuracy, which is determined by the chosen level of confidence. For 95% confidence intervals, the proportion of the computed confidence intervals that correctly contained the population parameter should be close to .95. Tong et al. (2012) defined accuracy as "the ability to measure the true value of the characteristic correctly on average" (p. 84).

The average interval widths represents an average length of n repeated confidence intervals. The length is computed by the difference between the upper bound and the lower bound of the interval (Tong et al., 2012). The mean width indicates how the interval is precise and informative in the sense of having small length. Kelly and Rausch (2006) argued that "holding the confidence interval coverage constant, the narrower the confidence interval, the more information about the population parameter of interest is obtained" (p. 381)

The coverage error represents the absolute difference between the nominal level (e.g., 95% CI) and actual coverage probability (Wu et al., 2003). This criterion identifies the probability that the interval does not cover the true value of the parameter, and the desired value for coverage error is zero (Hurairah, Akma Ibrahim, Bin Daud, & Haron, 2006).

The upper and lower error probability represent the percentage of a true parameter value falling above and below the intervals. This criterion identifies the symmetry of the confidence interval, and the desired value for symmetry of the upper and lower error probabilities depends on the selected confidence level (e.g., for 95% CI the error probability is .025). Few studies have paid attention to the balance between the left and right non-coverage rate (e.g., Newcombe, 1998; Swift, 2009). Hurairah et al. (2006) argues that the "error probabilities are symmetric when the large of the lower or upper error probability is less than 1.5 times the smaller one". However, "symmetry of error probabilities may not occur due to the skewness of the actual sampling distribution" (p. 141). Other studies argue that asymmetric confidence intervals should be preferred when the distribution is bounded and skewed (Cooley, 2013; Qin & Hotilovac, 2008).

Selection of an appropriate confidence interval method does not need to satisfy all the criteria simultaneously. Swift (2009) argued, "There is usually some tension between the different criteria." However, choosing the best interval "depends on the relative importance the investigator places on each of these principles" (p. 749). It is useful to use more than one criterion in evaluating the performance of different CI methods. Penfield (2003) argued that,

> If two methods yield nearly identical coverage rates but drastically different interval sizes, then the method providing the smaller interval size is to be preferred. Similarly, if two methods yield nearly identical coverage rates but different sizes of error for the trials

displaying noncoverage, then the method providing the smaller error associated with noncoverage trials is to be preferred. (p. 154)

CHAPTER 3: Method

The purpose of this study was threefold: First, to evaluate four alternative methods of describing the degree of uncertainty associated with estimates of the mean composite rating of instructor effectiveness averaged across the number of responding students in each class. Second, to examine how the degree of uncertainty in class means vary as a function of the number of responding students. Third, to examine how the dependability of mean SET ratings vary as a function of the number of responding students and number of items. The following areas are discussed in this chapter: (a) subjects, (b) instrument, (c) analysis of data, and (d) limitations.

Subjects

All classes with fewer than five students who responded to the SET questionnaire were excluded from analysis in this study. The total number of excluded classes was 1,060. Most of these consisted of (a) special topics and/or individualized study classes, (2) private instruction classes, and (c) thesis or dissertation classes. As a result the data analyzed in this included the responses of 26,543 students who completed the questionnaire for one or more classes during the Winter 2016 semester at BYU. Only the responses to the five questions intended to evaluate the instructor are included in this study. Responses to the five questions about achievement of the BYU Aims were not included in the analysis. The total number of classes included in the analysis was 3,953 and the total number of faculty who were rated was 1,930. The analyzed classes ranged in enrollment from 5 to 836. The average size of the 3,953 classes was 36.24 and the average response rate was 78.22%. The average number of responding students per class was 27.10 with a range from 5 to 639 (See Figure 3).

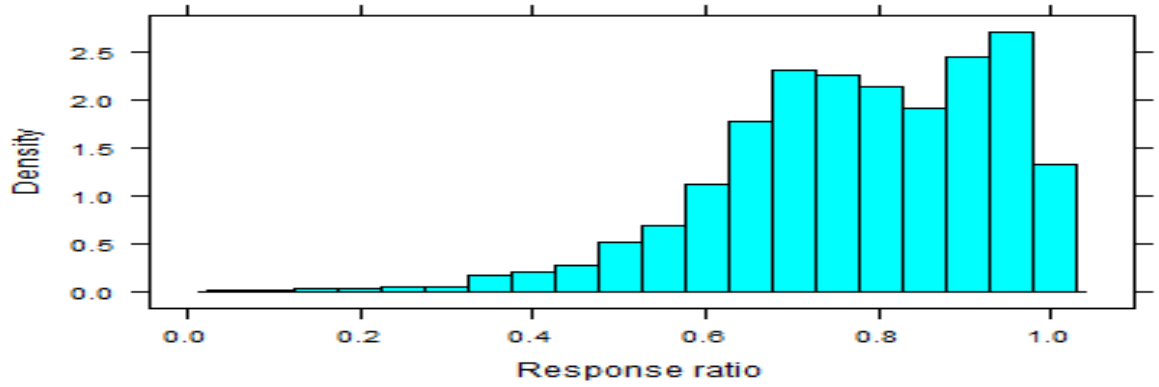

Figure 3. Distribution of response ratio for all class.

Instrument

The student ratings instrument used at Brigham Young University is designed to obtain feedback from students about teachers and courses as a means of helping faculty to improve their teaching. In addition, department chairs and university committees review and consider students' ratings as one source of information in making teacher retention and promotion decisions. This instrument consists of 10 questions and is administered every semester in each class.

A copy of this instrument is shown in Appendix A. All questions have 5-point scales, 1(Not at all effective) to 5 (Very effective). This analysis will focus only on the five items which related to teaching effectiveness. For example, one question asks:

How effective was this instructor (not the TA) in helping students who indicated a need for assistance?

1	2	3	4	5
Not at all effective	Not very effective	Moderately effective	Effective	Very effective

Analysis

Level of analysis. The class was used as the unit of analysis in this study. For the purpose of this study, a class is defined as a specific section of a particular course taught by a specific teacher at scheduled time during a particular semester. The class-level is based on a unique combination of variables used by BYU:

1. Semester_year (e.g., Winter 2016)
2. Course name (e.g., A HTG 100)
3. Instructor identification number (e.g., 1234567)
4. Class (e.g., 001002009045055). This variable consists of section numbers.

Based on the above combination the class id is (Winter 2016A HTG 100123456700100200904505 5).

Multilevel confirmatory factor analysis (MCFA). MCFA (Muthén, 1994) was used to evaluate the factor structure underlying the student rating scale. The decision was made to use a multilevel confirmatory factor analysis because of the hierarchical or nested structure of the data (Brown, 2014). Two levels were considered: class level and student level. The MCFA accounts for the within and between-level latent constructs by decomposing the total sample covariance matrix into pooled within and between-group covariance matrices and uses these two matrices in the analyses of the factor structure at each level (Brondino, Pasini, & da Silva, 2013; Dedrick & Greenbaum, 2011). Muthén (1994) argued that ignoring the hierarchical structure of the data would produce problematic results. More specifically, Dyer, Hanges, and Hall (2005) explained,

> when the total covariance matrix is factor analyzed, the fit of the group level factor structure as well as any factor loading estimates will be biased since it is a mixture of the factor structure operating at the between-group and within-group levels. Typically, this

total factor structure will primarily be a function of the within-group factor structure (p. 153).

As shown in Figure 4, the model consisted of single factor MCFA with five observed indicators ($X_1 - X_5$) for the five items in the scale. The five observed variables were hypothesized to load onto a single latent factor at the within level (F_W). The five item means ($\mu_1 - \mu_5$) were presumed to load onto the aggregated latent factor at the between level (F_B). Thus, "the observed values of the original indicators are considered to be a function of both the within- and between-level latent constructs" (Dyer et al., 2005, p. 154). The model also includes the within (λ_i^W) and the between (λ_i^B) factor loadings, and the within (θ_i^W) and between (θ_i^B) measurement error variances. The MCFA model was estimated using the robust Maximum Likelihood (MLR) estimator, which "has been found to be efficient in the estimation of latent variable models based on non-normally distributed responses and items rated on answer scales including five or more response categories" (Morin et al., 2014, p. 12).

Goodness of fit of the model was evaluated using the Tucker-Lewis Index (Bentler & Bonett, 1980); the Comparative Fit Index (Bentler, 1990) as relative fit indices; the Root Mean Square Error of Approximation (Hu & Bentler, 1999) as a parsimony corrected fit indices; Standardized Root Square Mean Residual (SRMR) as an absolute fit index. Values of.06 or less are considered an adequate fit for SRMR and RMSEA (MacCallum, Browne, & Sugawara, 1996). A value of .95 and above is considered an excellent fit for CFI and TLI. Adequacy of factor loadings and will be examined for the MCFA model. Factor loadings exceeding 0.40 are considered acceptable (Hair, Anderson, Tatham, & Black, 1995). The MCFA was conducted using Mplus version 7.4 (Muthén & Muthén, 2015).

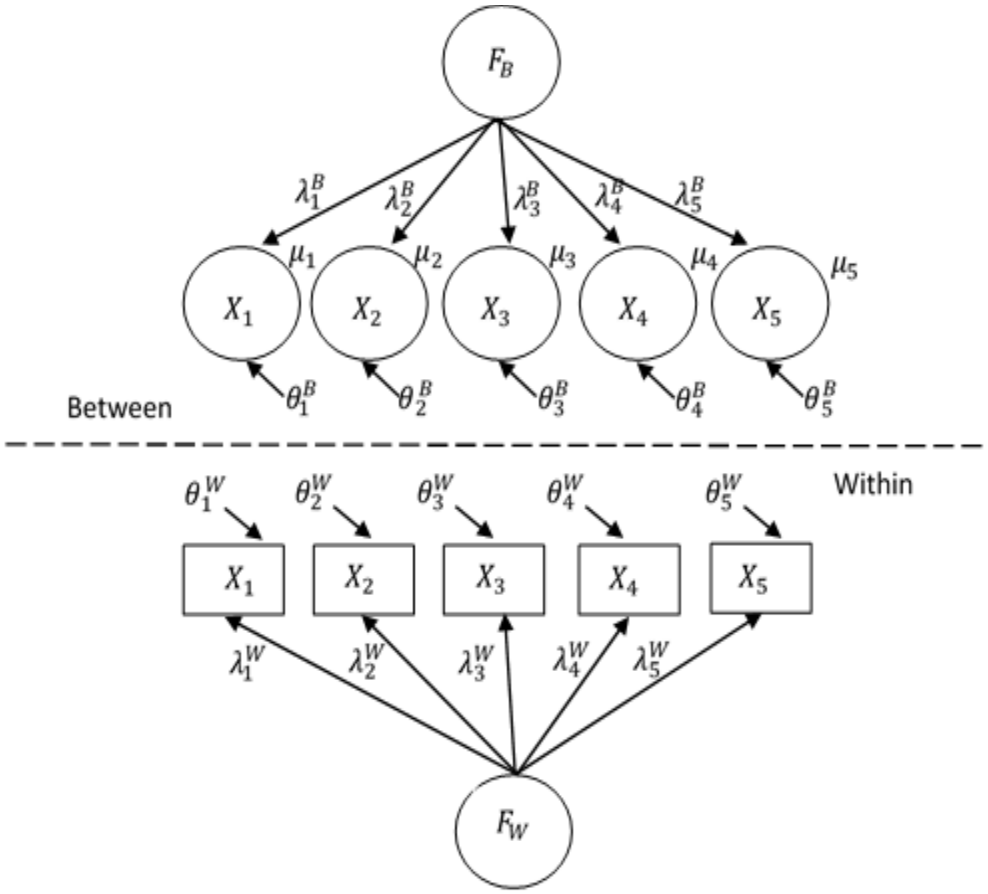

Figure 4. Path diagram of one-factor student ratings multilevel model.

Variance components and the estimated reliability coefficient. To answer the first research question, the result from MCFA was used to estimate the different variance components used in the generalizability coefficient equations.

Generalizability theory. Generalizability theory (Brennen, 2001; Cronbach et al., 1972) was used as the framework for estimating the reliability of the mean rating of the class level. Each student's responses to the five questions were averaged across items to obtain a student mean. The student means were then averaged across the responding students within each class to get the class mean. Variance components were then computed to estimate (a) the variance of classes, (b) the variance of classes by items, (c) the variance of students within classes, and (d)

the residual error variance. The variance due to item main effect and student mean effect will not included in this study because all classes are rated on the same set of items (items are crossed with classes), thus, any difference in items affect all classes and does not change their relative standing (Gillmore, Kane, & Naccarato, 1978; Webb, Shavelson, & Haertel, 2006). Generalizability theory of a student rating data structure is partially nested split plot design in which each student in each class is given the same set of items will response to the same items.

The design is $(i:p)$ X j (students (i) nested within class (p) crossed with items (j)). Where *i:p* is the students nested within class; *pj* is the items by classes variance; *ij:p* is the items by students nested within classes variance. In this study, students and items are considered important facets. Figure 5 illustrates the venn diagram of this design.

For the purpose of this study, multilevel confirmatory factor analysis was used to estimate the variance components for reliability estimation as shown in Table 1.

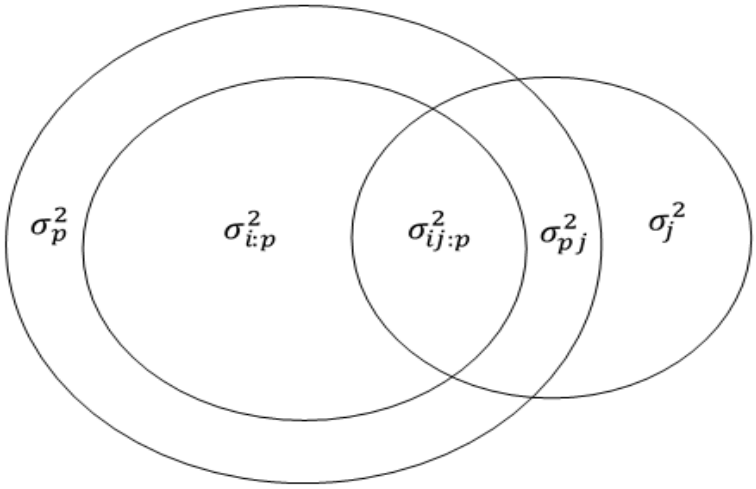

Figure 5. Venn diagram of i:p X j design.

Table 1

Variance Components

Between-classes	$\sigma_p^2 = \frac{\left(\sum_{i=1}^{j} \lambda_i^B\right)^2}{j^2}$	$\sigma_{pj}^2 = \frac{\left(\sum_{i=1}^{j} \theta_i^B\right)}{j}$
Within-classes	$\sigma_{i:p}^2 = \frac{\left(\sum_{i=1}^{j} \lambda_i^W\right)^2}{j^2}$	$\sigma_e^2 = \frac{\left(\sum_{i=1}^{j} \theta_i^W\right)}{j}$

Note. The σ_p^2 is the variance of classes; σ_{pj}^2 is the variance of classes by items; $\sigma_{i:p}^2$ is the variance of students within classes; and σ_e^2 is residual; *j* is the number of items; λ_i^W is the within factor loadings; λ_i^B is the between factor loadings; θ_i^W is the within measurement error variances; and θ_i^B is the between measurement error variances.

Two different conceptualization of the universe of generalization and two different formulas to estimate the reliability of class-level scores were considered (Kane & Brennan, 1977; Schweig, 2013). First, items were treated as a random (equation 3.1) where the selected items are drawn from and are intended to represent, a broader universe of such items (Kane & Brennan, 1977). Second, items were treated as a fixed (equation 3.2) where the selected items exhaustively define the universe to which generalization and inference are intended (Kane & Brennan, 1977). In both approaches, the students are treated as random.

$$E\rho_g^2 = \frac{\sigma_p^2}{\sigma_p^2 + \frac{\sigma_{pj}^2}{n_j} + \frac{\sigma_{i:p}^2}{n_i} + \frac{\sigma_e^2}{n_i n_j}} \qquad (3.1)$$

$$E\rho_g^2 = \frac{(\sigma_p^2 + \frac{\sigma_{pj}^2}{n_j})}{(\sigma_p^2 + \frac{\sigma_{pj}^2}{n_j}) + \frac{\sigma_{i:p}^2}{n_i} + \frac{\sigma_e^2}{n_i n_j}} \quad (3.2)$$

In (3.1) and (3.2) the $E\rho_g^2$ is the between-classes generalizability coefficient; the σ_p^2 is the variance of classes; σ_{pj}^2 is the variance component for the class by items interaction; $\sigma_{i:p}^2$ is the variance component for students nested within classes; and σ_e^2 is the residual variance. n_j the number of items; and n_i is average number of respondents.

Then, a D-study was conducted to determine an efficient number of items and students from the existing SET data required to obtain small error variances and/or large reliability coefficients which in turn will be used in making different evaluation decisions.

Intraclass correlations coefficient. ICC_1 and ICC_2 were computed to determine whether aggregated individual-level ratings are reliable indicators of group-level constructs. Both indices utilize one-way random-effect analysis of variance, having student rating at Level 1 as the dependent variable and grouping level (class) is the independent variable (Lüdtke et al., 2009). ICC_1 describes the amount of variance in each item that can be attributed to belonging to the class-level. In contrast, ICC_2 provides a reliable estimate of the class-level group means (Woehr et al., 2015).

Uncertainty in estimation of the universe mean for a class. To answer the second and third research questions, five methods were used to construct the confidence interval of class mean as follows:

Logit transformation method. This analysis was carried out under R, version 3.2.2. (R Core Team, 2015. The logit transformation method was used to analyze student ratings which

will make the data more closely conform to theory and to a normal distribution and thus easier to compute the confidence interval (Williamson & Gaston, 1999). Some researchers argue that the logit transformation should be used when the data are bounded interval (Gart et al, 1985; Lesaffre et al., 2007). Let $x = \{x_1, \ldots, x_n\}$, then the procedure to find the 95% confidence interval of the mean, is:

- The scale score *x* ranging from 1 to 5 will be transformed into scale score *x* ranging from 0 to 4 by subtracting 1 from each ($x = x - 1$).
- The transformed scale score of *x* will be transformed into a Proportion of Maximum Possible (POMP) score *p* ranging from 0 to 1 by dividing it by 4 ($p = x/4$).
- A rescaled ("shrunken") probability $\acute{p}$ is mapped to the observed probability *p* by the following,

$$\acute{p} = \delta p + .5(1 - \delta). \quad (3.3)$$

Here, δ is a scaling constant greater than .5 and less than 1. We will choose δ to be .95. Warton and Hui (2011) argued that,

> one difficulty, though, with using this transform is that sample proportions equal to 0 and 1 transform to undefined values $-\infty$ and ∞, respectively. An ad hoc solution to this problem is to add some small value ε to both the numerator and denominator of the logit function (p. 5).

- The modified logit for the rescaled probability is

$$t = \ln(\frac{\acute{p}}{1 - \acute{p}}) \quad (3.4)$$

- Estimate the mean, SD, confidence interval from the *t* values.
- A modified logit *t* will be transformed back to the rescaled probability $\acute{p}$ using the following,

$$\hat{\acute{p}} = \frac{e^t}{1 + e^t} \tag{3.5}$$

- The rescaled probability $\hat{\acute{p}}$ will be transformed back to estimated standard probability $\hat{p}$ with the following,

$$\hat{p} = \frac{\hat{p} + .5\acute{(}\delta - 1)}{\delta} \tag{3.6}$$

- The POMP probability $\hat{p}$ will be transformed back into the rescaled value of a variable $\hat{x}$ score from 0 to 4 ($\hat{x} = 4\hat{p}$).
- The estimated score will be transformed from the 0 to 4 scale back to a 1 to 5 scale by adding 1 ($\hat{X} = \hat{x} + 1$).

Resampling method. Resampling method or percentile confidence interval for a population mean method (Karian & Dudewicz, 2011). This analysis was carried out under R, version 3.2.2. Let $x = \{x_1, \dots, x_n\}$, the procedure to find the 95% confidence interval of the mean, following (Wilcox, 2001), is:

- Independently generate *1000* random samples of size n, with replacement.
- For each sample, estimate the mean $\hat{\theta}$.
- Obtain *1000* estimates $\hat{\theta}_1, \hat{\theta}_2, \dots , \hat{\theta}_{1000}$.
- Place the *1000* estimates $\hat{\theta}_1, \hat{\theta}_2, \dots , \hat{\theta}_{1000}$ in increasing numerical order, obtaining

$$\hat{\theta}_{(1)} \leq \hat{\theta}_{(2)} \leq \dots \leq \hat{\theta}_{(1000)}$$

- The percentile method's $100(1 - \alpha)\%$ confidence interval for θ is

$$(\hat{\theta}_{(a)}, \hat{\theta}_{(b)})$$

where

$$a = \left\lfloor 1000 * \frac{\alpha}{2} \right\rfloor, b = \left\lceil 1000 \left(1 - \frac{\alpha}{2}\right) \right\rceil$$

This indicates selecting the 25th and 975th values to construct 95% confidence interval.

Non-parametric bootstrapping (Accelerated Bias-Corrected BCa). This analysis was carried out with the bootstrap library (Tibshirani & Leisch, 2015) under R, version 3.2.2 (R Core Team, 2015). In this analysis, we used the accelerated bias-corrected percentile limit (BCa) which performed a resampling with replacement. Let $x = \{x_1, \dots, x_n\}$, a sample of size n. To approximate the sampling distribution of θ, we took 1000 resamples from x and use these samples for computing the confidence interval. Then, following Kelley (2005), the function computed the bias correction value $\hat{z}_0$ by calculating the proportion of the bootstrap distribution d^* values that are less than the sample x, then found the 25th and 975th from the normal distribution with that cumulative probability:

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\#\,(d^* < x)}{n}\right) \tag{3.7}$$

In (3.7) Φ^{-1} is the inverse cumulative distribution function for the standard normal distribution. The acceleration constant, $\hat{a}$, can be computed using:

$$\hat{a} = \frac{\sum_{i=1}^{n}(\tilde{d} - x_{-i})^3}{6\left(\left(\sum_{i=1}^{n}\left(\tilde{d} - x_{-i}\right)^2\right)^{\frac{3}{2}}\right)} \tag{3.8}$$

In (3.8) $\tilde{d}$ is the mean of *n* jackknife x_{-i} values. Then we computed the confidence intervals from the bootstrap sample using:

$$Lower = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{\left(\frac{\alpha}{2}\right)}}{1 - \hat{a}\left(\hat{z}_0 + z^{\left(\frac{\alpha}{2}\right)}\right)}\right) \tag{3.9}$$

$$Upper = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-\frac{\alpha}{2})}}{1 - \hat{a}(\hat{z}_0 + z^{\left(1-\frac{\alpha}{2}\right)})}\right) \quad (3.10)$$

Bayesboot method. This analysis was carried out with the Bayesboot library (Baath, 2016) under R, version 3.2.2 (R Core Team, 2015). This method performs the Bayesian bootstrap introduced by Rubin (1981). Two steps are needed to perform this method as described by Baath (2016). Let $x = \{x_1, \ldots, x_n\}$, to generate a Bayesian bootstrap sample of size n, repeat the following n times:

1. Draw weights from a uniform Dirichlet distribution with the same dimension as the number of data points. The Dirichlet distribution is a multivariate distribution over proportions of real numbers between 0.0 and 1.0 that together sums to 1.0, and where any combination of values is equally likely. O'Hagan and Stevens (2003, p. 47) argued that the "key result is that we can draw random population distributions from this posterior distribution in a very similar way to the bootstrap."
2. Calculate the statistic (e.g., mean) of each of these Bayesian bootstrap samples, using the Dirichlet, and record it.

For the purpose of estimating the class mean and the range of uncertainty, we did a Bayesian bootstrap analysis of logit transformation procedure mentioned before by creating a function that takes the data and returns the mean, lower bound, and upper bound. Bayesboot procedure sampled from the data according to the probabilities defined by the Dirichlet, and use this resampled data to calculate the statistic. We used 1000 as the number of bootstrap draws and class size as the size of the original data used to calculate the statistical parameter for each Dirichlet draw.

Classical Z distribution method. This analysis was carried out using R, version 3.2.2 (R Core Team, 2015). Classical Z method is the most common method used to compute confidence interval of SET means. Estimation of the margin of error using the classical Z method is based on the central limit theorem (CLT) which assumes that the population of class means is normally distributed and the resulting interval is symmetrical about the estimated mean. Because the distribution of class means in the context of this study was known to be markedly skewed, we purposefully explored other options that do not depend on the normality assumption. However, the classical Z method was used as a benchmark to compare its results to the other four non-traditional methods in this study. Let $x = \{x_1, \ldots, x_n\}$, then the procedure to find the 95% confidence interval of the mean, is:

- Estimate the mean and the standard deviation of each class.
- Estimate the standard error by dividing the standard deviation by the square root of the class size.
- Estimate the margin of error; for a 95% confidence level the margin of error is 1.96 times the standard error.
- Estimate the lower limit (mean - margin of error) and the upper limit (mean + margin of error) of the confidence interval.

Simulation study. Since a theoretical comparison of the estimators is not possible, the researcher carried out a simulation study to compare the performance of the confidence interval with five methods: the logit transformation, the resampling, the bootstrapping BCa, the Bayesian bootstrapping, and classical Z method in finite sample sizes. The simulation procedure following (Shi & Golam Kibria, 2007, p. 416) is:

1. Select the sample size (n), number of simulation times (1,000), and confidence level (e.g., 95%).
2. Generate $x_1, x_2, \ldots, x_n$ simulated student ratings which represent an independent and identically distributed (iid) sample from a beta distribution with two parameters α and β.
3. Construct confidence intervals for the population mean using each of the five methods.
4. For each method in step 3, determine if the confidence interval includes the population mean, and calculate the width of the interval, and the upper and lower error.
5. Repeat (1)–(4) 1,000 times, then compute the coverage probability, the average of the width, the coverage error and upper and lower error probability.

We considered five sample sizes in the simulation study: n = 5, 10, 20, 50, 100. This process used 1,000 simulated random samples bounded between 1 and 5 with various combinations of true means, standard deviation, and sample sizes that can correspond the real-life situations of interest. Moreover, every sample is resampled 1,000 times for the resampling, Bootstrapping, and Bayesian boot methods. The simulations, including all data generation, is programmed in R, version 3.2.2 (R Core Team, 2015).

Data generation. The research used a conditional beta distribution as a parametric model to create the different shapes of distributions (Nelson & Preckel, 1989). The beta distribution is bounded between 0 and 1 and is characterized by two shape parameters, α, and β. Moreover, it has the flexibility to transform a rich family of distributional shapes into the standard beta distribution when a distribution over some finite interval is needed (Cordeiro & de Castro, 2011; Farnum & Stanton, 1987). The density function following (Kong, Parker, & Sul, 2014, p. 10) is given by

$$f(x|\alpha,\beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad \text{for } x \in [0,1], \alpha > 0, \beta > 0, \tag{3.11}$$

In (3.11) $B(\alpha, \beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$, which is the beta function.

To generate random data bounded between lower bound (Lb) and upper bound (Ub) that has a specific distributional shape, different combinations of the mean θ and standard deviation σ are chosen. The values of the mean θ and standard deviation σ used to compute the beta distribution parameters needed to generate the data. For the first combination $\theta = 2.9, 3.0, 3.1$ which represents symmetric distributions, the second combination has $\theta = 4.2, 4.4, 4.7$ which represents left-skewed distributions, and the last combination has $\theta = 1.5, 1.8, 2.0$ which represents right-skewed distributions. Three values of $\sigma = 0.05, 0.10, 0.20$ used with each combination. These small values were selected because of the observed homogeneity in the SET ratings (Appendix B). The procedure used for data generations following (AbouRizk, Halpin, & Wilson, 1991), is:

1. Determine the desired shape by specifying the mean θ and standard deviation σ.
2. Specify the sample size n.
3. Compute $\alpha = \left[\left(\frac{\theta - Lb}{Ub - Lb}\right)\left(\frac{(\theta - Lb)(Ub - \theta)}{s^2}\right) - 1\right]$ (3.12)
4. Compute $\beta = \left(\frac{Ub - \theta}{\theta - Lb}\right)\alpha$ (3.13)
5. Generate the random beta data using *n*, α, and β using *rbeta* function under R, version 3.2.2.
6. Transform back the data to original bounds: $Lb + data(Ub - Lb)$.

Measures of performance. The performance of each method in the simulation study was judged by the following criteria:

1. Coverage probability: For each simulated combination, we calculated the confidence intervals for the estimate mean for all the methods. The coverage probability for each method is calculated as the number of intervals containing the true mean divided by 1000.
2. Average interval width: For each simulated combination, we calculated the average width of each method. The width of the 95% confidence interval is the average difference between the upper limit and lower limit.
3. Coverage error: For each simulated combination, we calculated the absolute difference between the nominal level (e.g., 95% CI) and actual coverage probability.
4. Upper and lower error probability: For each simulated combination, we calculated the percentage of a true parameter value falling above and below the intervals.

CHAPTER 4: Results

Demographics, Instrument, and Descriptive Statistics

As mentioned in Chapter 3 there were 26,543 students in the sample who completed the questionnaire for one or more classes during the Winter 2016 semester at BYU. The total number of classes included in the analysis was 3,953 and the total number of faculty who were rated was 1,930. The analyzed classes ranged in enrollment from 5 to 836. Table 2 shows the number of classes in different enrolment ranges with their percentage. On one hand, the majority of classes have enrollment range of 11-20 students (30.4%). On other hand, the results show that more than 50% of classes have enrolment of 21 students or more. Out of 3,953 classes analyzed in this study 3,080 classes have a response rate higher than 61% as shown in Table 3.

Table 2

Number of Classes per Enrollment Range

Enrolment range	5-10	11-20	21-30	31-40	41-50	51-100	101-900
Number of classes	524	1200	831	425	284	470	219
Enrolment %	13.3	30.4	21.0	10.8	7.2	11.9	5.5

Table 3

Number of Classes per Response Ratio Range

Response ratio range	0-20 %	21-40 %	41-60 %	61-80 %	81-100 %
Number of classes	6	160	707	1541	1539

Descriptive statistics for the items and scales are reported in Tables 4 and 5. Item means ranged from 4.312 (SD = 0.896) for the first item to 4.598 (SD= 0.730) for the last item. There

were no missing values in this data set. The distribution of responses to each item was left skewed distributed, with skewness ranging from -1.308 to -2.122 and kurtosis values ranging from 1.344 to 4.978.

Table 4

Distribution of Student Responses by Item

Item	Not at all effective	Not very effective	Moderately effective	Effective	Very effective
1	1.089 %	3.468 %	12.422 %	29.152 %	53.869 %
2	1.333 %	3.684 %	11.820 %	27.698 %	55.466 %
3	0.662 %	2.205 %	9.673 %	27.526 %	59.934 %
4	0.653 %	2.494 %	10.087 %	25.698 %	61.068 %
5	0.685 %	1.494 %	6.045 %	20.901 %	70.875 %

Note. N= 3953.

Table 5

Item Descriptive for the SET

Item	Item description	Mean	SD	Skewness	Kurtosis	ICC
1	Helping students	4.312	0.896	-1.308	1.344	.201
2	Providing opportunities	4.323	0.914	-1.388	1.548	.198
3	Teaching challenging concepts	4.439	0.807	-1.510	2.159	.191
4	Demonstrating respect	4.440	0.822	-1.519	2.037	.183
5	Organizing the course content	4.598	0.730	-2.122	4.978	.198

Note. N= 3953.

Data Analysis and Results of the Research Questions

Dependability of student ratings. The first research question was, "what is the reliability of the estimated class means for the composite instructor effectiveness ratings?"

The reliability of the SET ratings was estimated by computing generalizability coefficients. As mentioned in Chapter 3, this study illustrates a two-facet partially nested split-plot design. For this design $(i{:}\,p)$ X j, four variance components were estimated: (a) the variance of the classes, (b) the variance of the classes-by-items interaction, (c) the variance of students within classes, and (d) the residual error variance.

A Multilevel Confirmatory Factor Analysis (MCFA) model was used to estimate the anticipated sources of variation in the SET data. Prior to conducting the MCFA, an intraclass correlation coefficient (ICC) was calculated for each of the five items in the SET to estimate the variability between and within classes on each item and the degree of non-independence or clustering in the ratings. Table 5 displays the ICCs for each of the five items. The ICCs ranged from .183 for item 4 to .201 for item 1. These values indicate that there is sufficient between class variability to warrant multilevel analysis. The estimated reliability for the class-level mean using ICC_2 in this study, with an average cluster size of 27 respondents per class, was .867. However, reliability is a variable rather than a constant in the context of SET ratings. It will be less than this estimated value for classes with fewer respondents and larger for classes with a greater number of respondents.

Table 6 shows the standardized factor loadings for this model. All factor loadings were significantly greater than zero ($p < .01$) and adequate (all greater than .64 for the standardized solution) suggesting that all five items adequately reflect the latent construct. The standardized loadings for the items load strongly onto the single factor at the between level, ranging from .853

to .965. The factor loadings of the items at the within level, ranging from .691 to .780. Between-level loadings were stronger than within level, underlining the importance of the group level for SET data.

The fit indices suggest that the data-model fit was acceptable. The Root Mean Square Error of Approximation (RMSEA) = .075, CFI = .977, and TLI = .952. The Standardized Root Mean Square Residual (SRMR) fit indices at each level indicated adequate fit (SRMR-within = .024 and SRMR-between = .035). The path diagram of the model is displayed in Figure 6.

Table 6

Unstandardized and Standardized Factor Loadings by Item

	Unstandardized		Standardized	
Item	Within	Between	Within	Between
1	0.641	0.375	0.797	0.930
2	0.617	0.366	0.745	0.887
3	0.568	0.343	0.775	0.965
4	0.582	0.329	0.780	0.928
5	0.460	0.282	0.691	0.853

Table 7 presents the estimates of the four sources of variation including (a) estimated variance components, (b) the total variance, and (c) percentage of the total variation. The results showed that students-within classes represents the largest variance component and that it accounts for 47% of the total variance. This finding indicates that the students' ratings of classes were greatly affected by differences in the ratings assigned by the students. The second largest variance component accounted for 16% of the total variance and represents class-to-class variability.

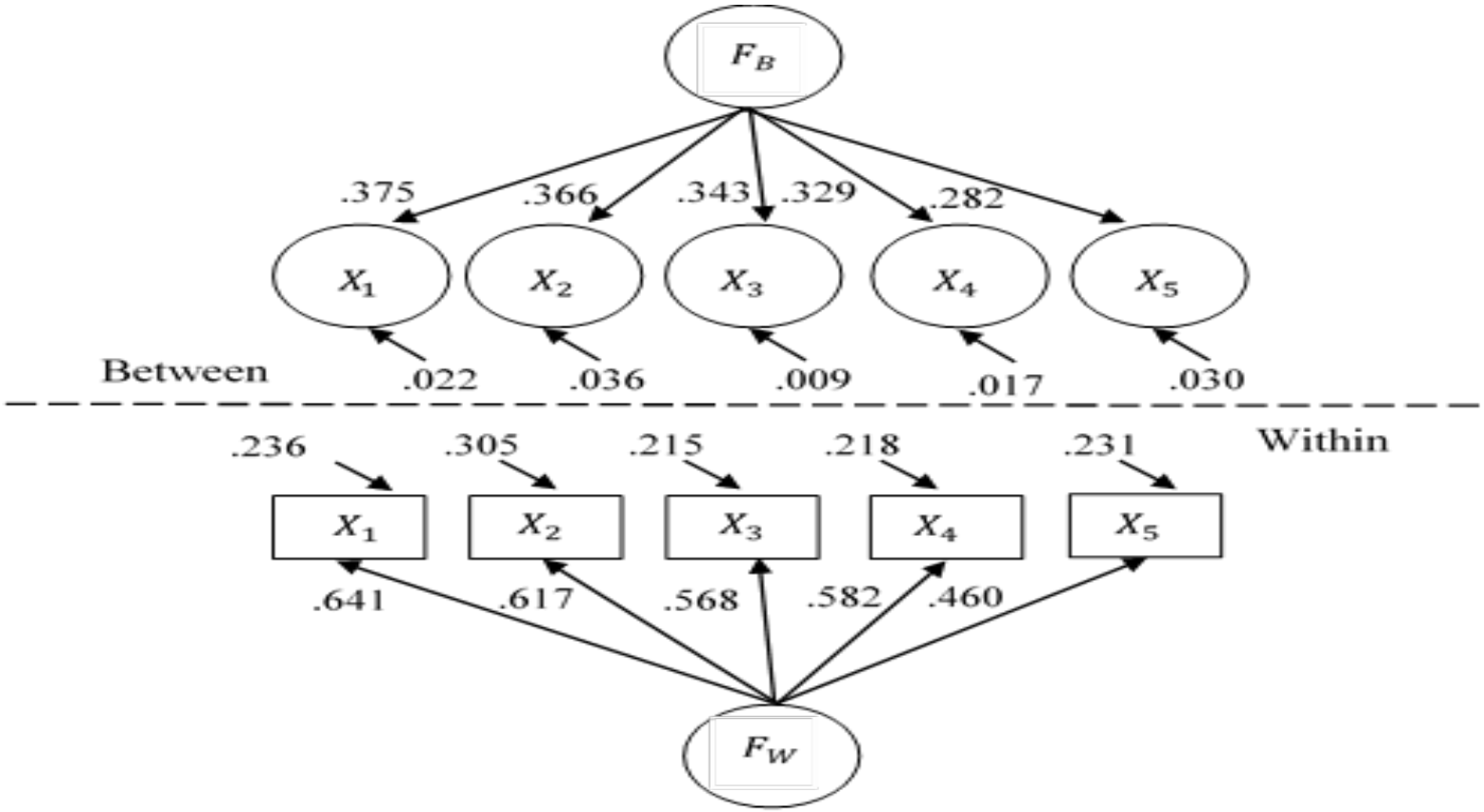

Figure 6. Path diagram and unstandardized parameters estimates of the multilevel model for the SET.

The smallest variance component represents the variability due to class-by-item interaction which accounted for 3%. The results show that with the average of 27 respondents, the generalizability coefficients when items were treated as a fixed facet and when items were treated as a random facet were .895 and .861 respectively.

Table 7

Estimated Variance Components

Source of Variance	Estimated Variance Component	Percentage of Total
Classes (σ_p^2)	0.115	16%
Classes by Items (σ_{pj}^2)	0.023	3%
Students within classes ($\sigma_{i:p}^2$)	0.330	47%
Residual (σ_e^2)	0.241	34%
Total variance	0.709	100%

Figure 7 shows how the generalizability coefficients are expected to vary as a function of the number of items. The results show that high reliability is obtained when items are treated as a fixed facet rather than as a random facet. The G coefficients are not greatly influenced by changing the number of items in either condition except when only one item used in the random condition. In general, reliable results can be achieved using different number items, with the average of 27 students responds to SET questionnaire. On the other hand, Figure 8 shows that the G coefficients are influenced by varying the number of students who respond to the five items. Highly reliable results obtained when the average number of students were 15 for fixed item design and 27 for random item design. The key point here is that the reliability of the class means is a variable rather than a constant. It varies as a function of the number of respondents in a class. The greater the number of respondents the larger the reliability.

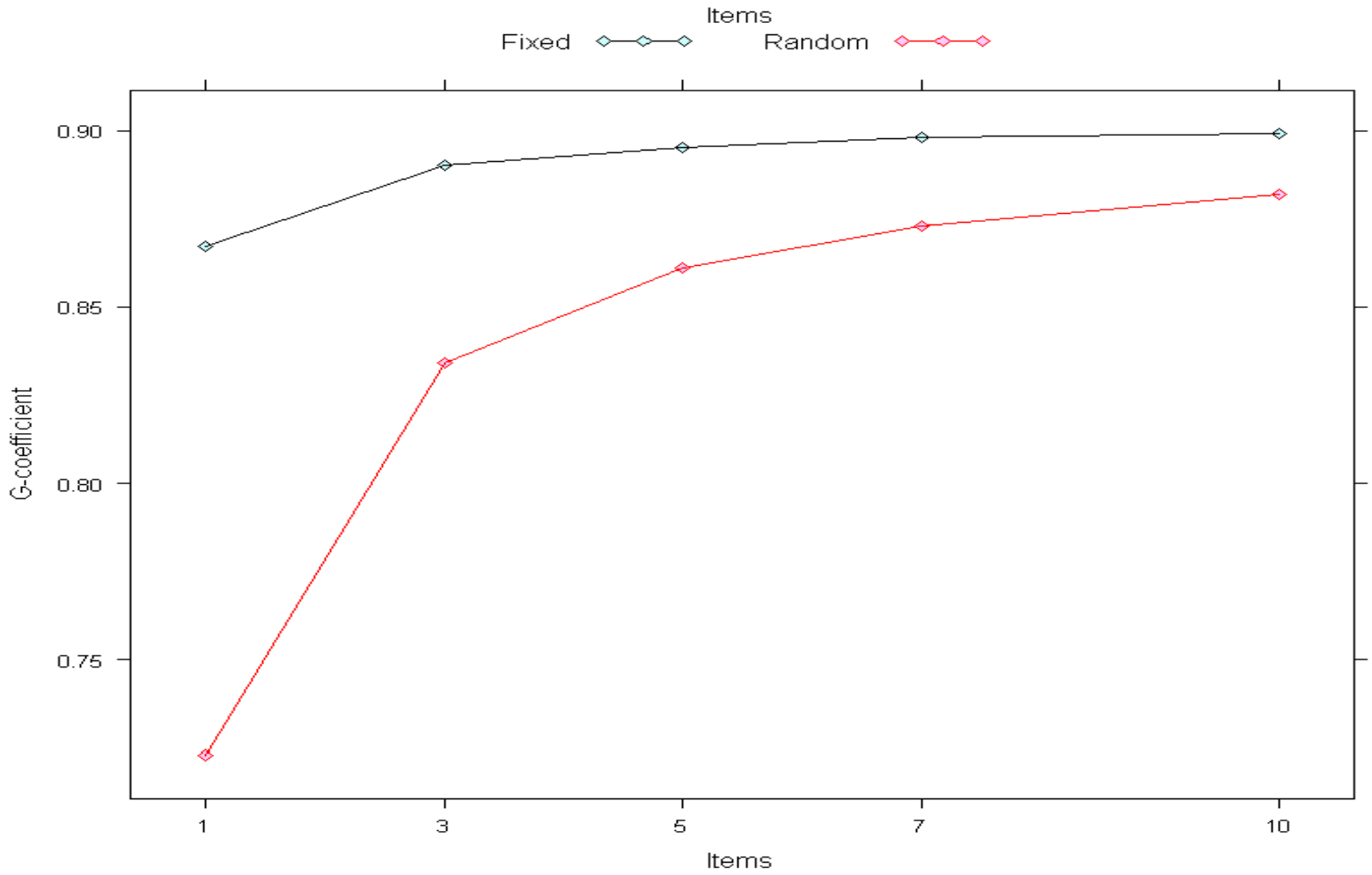

Figure 7. Generalizability coefficients for various numbers of items.

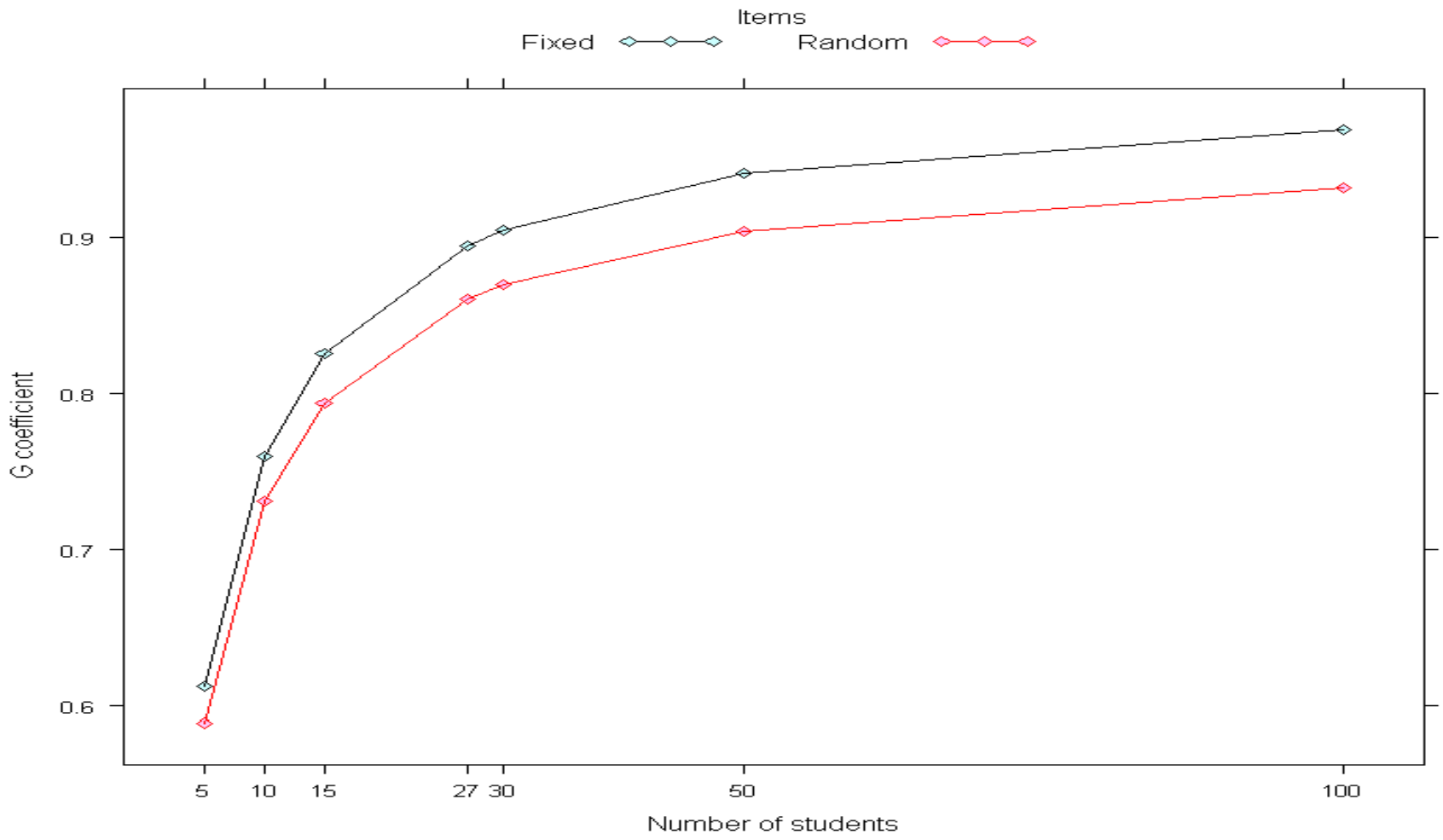

Figure 8. Generalizability coefficients for various numbers of respondents.

Figure 9 shows the generalizability coefficients when varying both the number of items and the average number of respondents. When items are treated as a fixed facet, highly reliable results cannot be achieved regardless of the number of items used when the number of respondents is 10 or less. The average number of respondents required to obtain highly reliable results is 15 or more with at least 3 items. The results show that the average number of respondents has a positive effect on the dependability of the ratings. Large number of respondents give more reliable ratings and small number of respondents give less reliable ratings. When items are treated as a random facet, highly reliable results achieved when three or more items used and the average number of respondents required is 27 or more.

Uncertainty in estimation of the universe mean for a class. Our main objective in this study is to find the best interval estimator for describing the degree of uncertainty associated with estimates of the mean composite rating of instructor effectiveness averaged across the number of responding students in each class. A simulation study was conducted in hopes of accomplishing this objective since a theoretical comparison is difficult. The three families of distributions considered include the (a) symmetrical, (b) right-skewed, and (c) left-skewed families that correspond to the real-life situations of interest (see Figures B1- B3) in Appendix B. The primary outcomes of interest are: (a) the coverage probabilities of the 95% confidence interval estimates, (b) the coverage error probabilities, (c) the average interval width, and (d) the degree and direction of asymmetry of the interval about the mean. In addition to the four methods of interest in this study, we included a classical method (e.g., Z method) for comparison purposes.

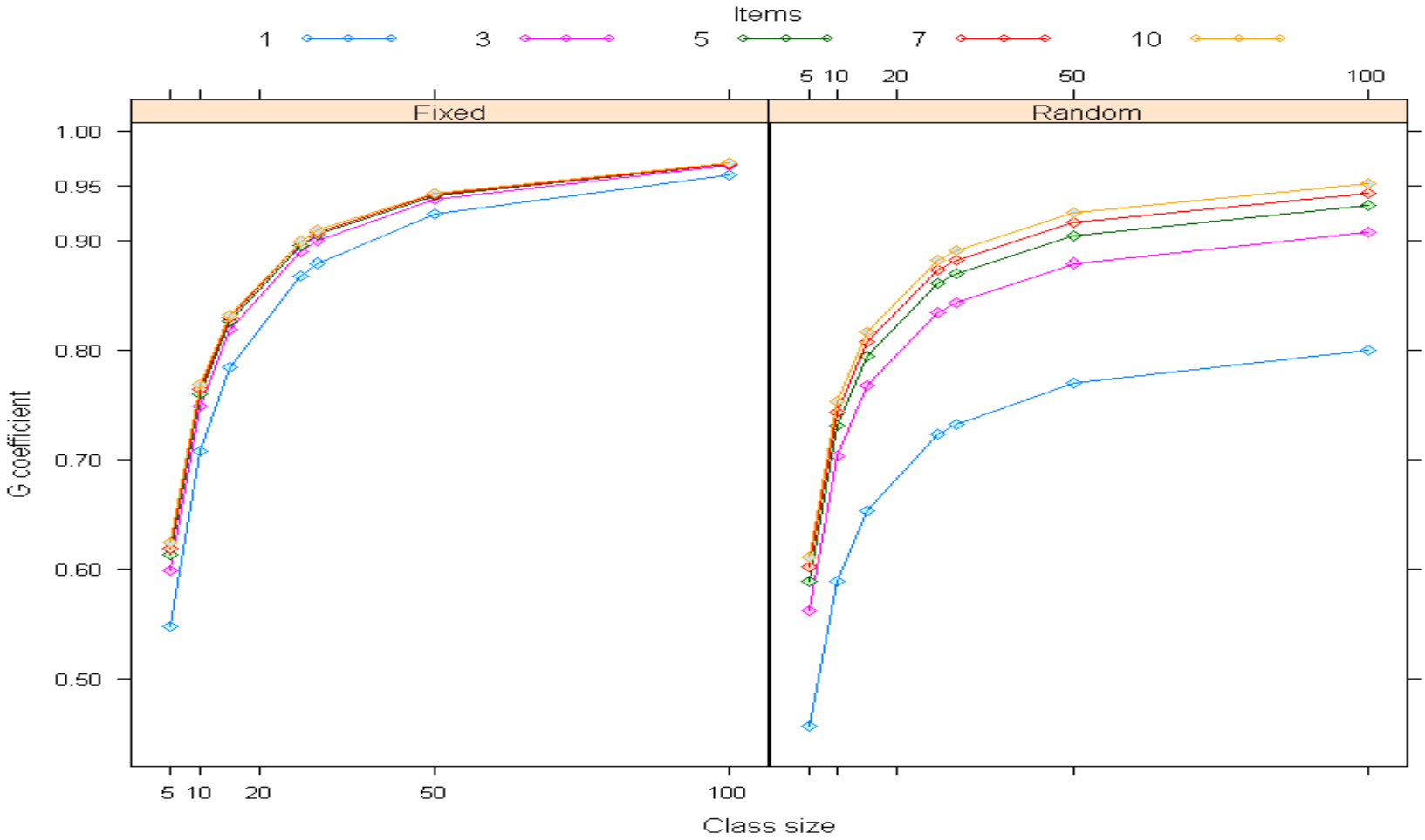

Figure 9. Generalizability coefficients for various number of items and various number of respondents.

Symmetric distributions. For comparison convenience, the results of the simulation from symmetric distributions are presented here in plot form and reported in tables (see Tables D1-D9) in Appendix D.

Coverage probabilities. The results in Figure 10 suggest that when sampling from a symmetric distribution, the logit transformation method has coverage probability close to the nominal level of 95% and remains the same for different sample sizes. However, for small sample sizes, the estimated coverage probability of all methods (except the logit transformation) are below the nominal level of 95%. As the sample size increases, the estimated coverage probability of these methods increases. However, it is obvious that the Z distribution method outperformed the other bootstrap methods with sample sizes of 10 or more, because the coverage probabilities with this method tend to be higher than those with the other bootstrap methods. The Resampling, BCA, and Bayesboot need a sample size of 50 or more to attain a value close to the nominal level of 95%.

Coverage error probabilities. The results for the coverage error probabilities are shown in Figure 11. These results are associated with the results obtained from the coverage probabilities criterion. The logit transformation method has almost zero coverage error for different sample sizes. We observe that the coverage error probabilities for the other methods are relatively large for small sample size. However, as the sample size increases, the estimated coverage error probability of these methods decreases.

Average width. The results for the average interval width are shown in Figure 12. We observe that the interval width of all methods depend upon both the sample size and the magnitude of the variance in the underlying distributions. As the variance of the sample increases, the estimated interval width increases, and as the sample size decreases, the estimated

interval width increases. All methods produced the same interval width with sample size of 20 or more. Moreover, we observe that the resulting interval of the logit method was marginally wider on average than the other methods when the sample size is 10 or less.

Upper and lower error probabilities. The results for the upper and lower error probabilities when sampling from a symmetric distribution for the logit method are shown in Figure 13. The results show that the estimated confidence intervals of the logit method are symmetrical for different sample sizes. However, the other methods require large sample size to obtain symmetrical confidence intervals (See Figures E1, E4, E7, & E10) in Appendix E.

Right-skewed distributions. In this section overall coverage probabilities, coverage errors, average widths, and upper and lower error probabilities of the resulting confidence intervals estimated by each method considered are given for the right-skewed distributions. The results are presented here in plot form and reported in tables (see tables D10-D18) in Appendix D.

Coverage probabilities. The results in Figure 14 suggest that when sampling from a right-skewed distribution, the logit transformation method has coverage probability close to the nominal 95% level and remains the same for different sample sizes. However, for small sample sizes, the estimated coverage probability of all other methods are below the nominal level 95%. As the sample size increases, the estimated coverage probability of these methods increases.

However, it is obvious that the Z distribution method outperformed the other bootstrap methods with sample size of 20 or more, because the coverage probabilities with this method tended to be higher than those with the other bootstrap methods. The resampling, BCa, and bayesboot need a sample size of 100 or more to attain a value close to the nominal level 95%.

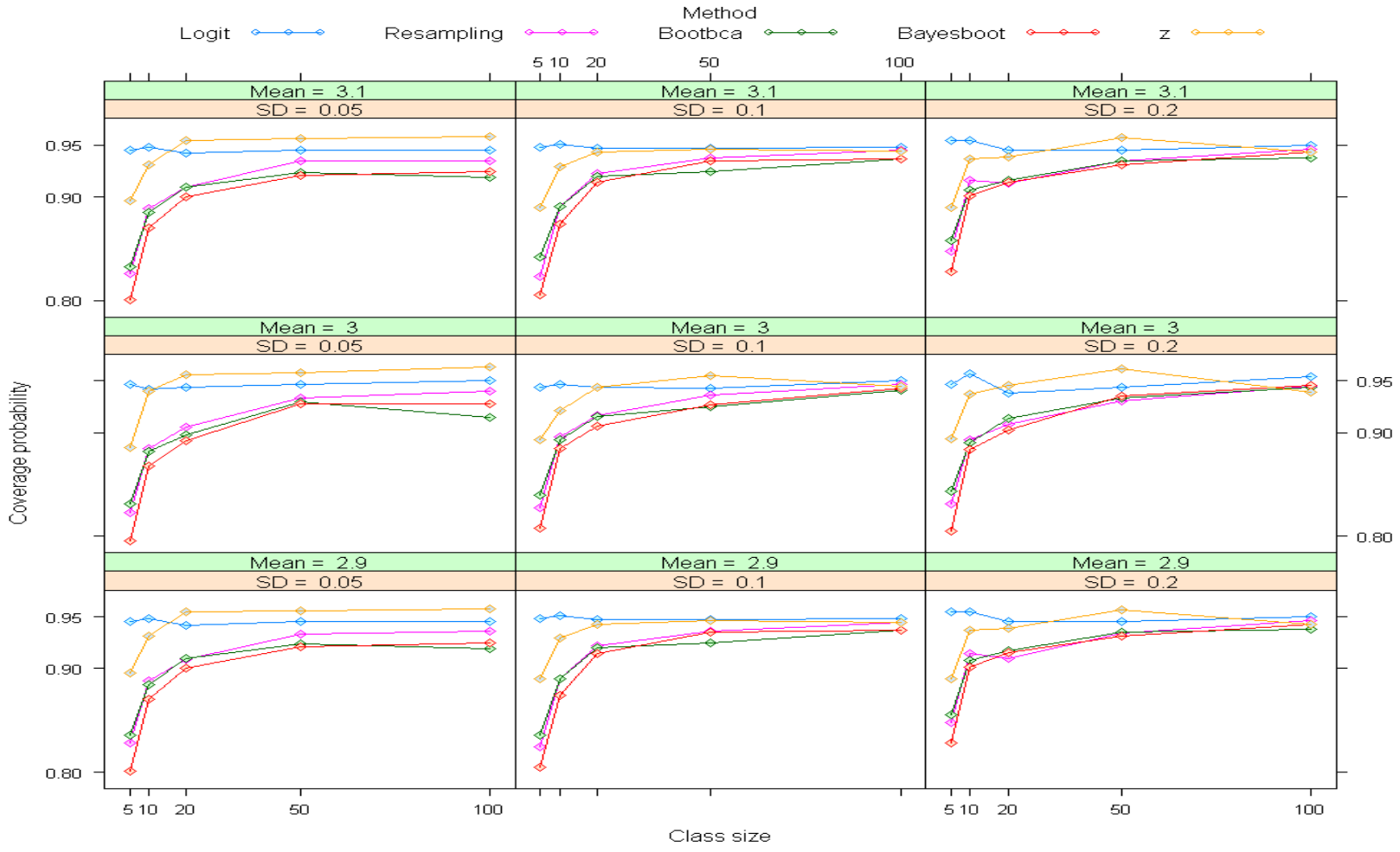

Figure 10. Coverage probabilities of the 95% CIs for the symmetrical distributions.

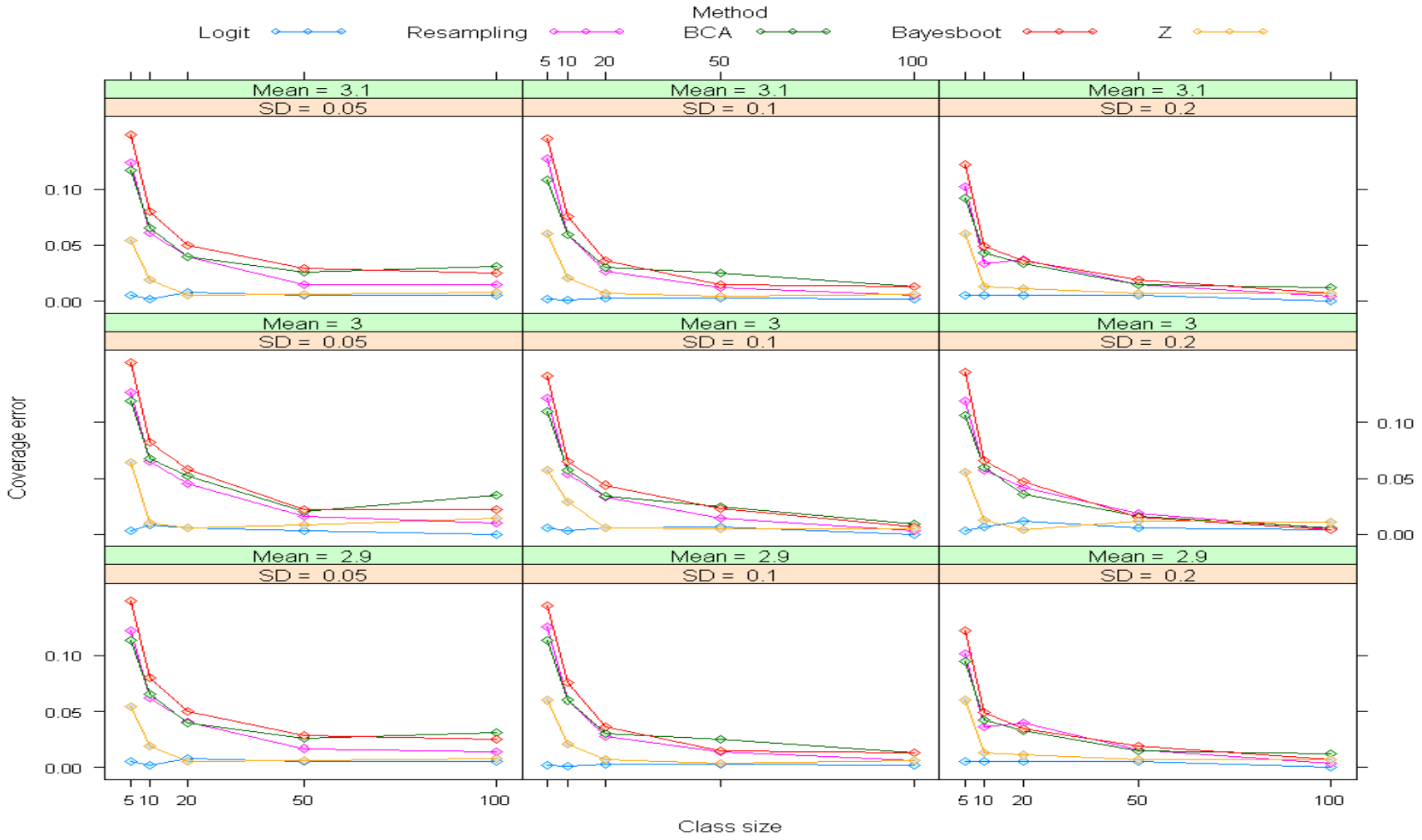

Figure 11. Coverage error probabilities of the 95% CIs for the symmetrical distributions.

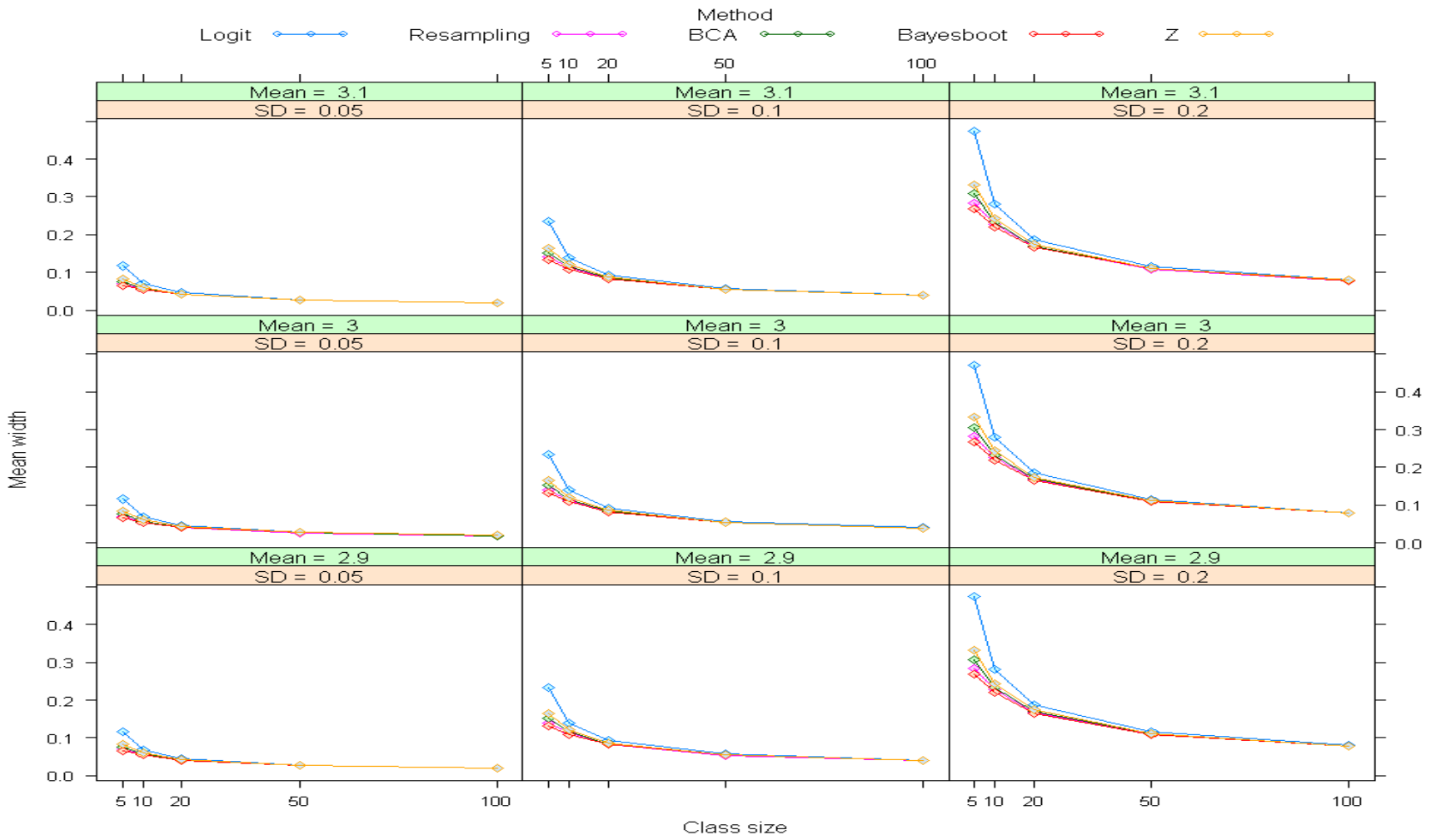

Figure 12. Average width of the 95% CIs for the symmetrical distributions.

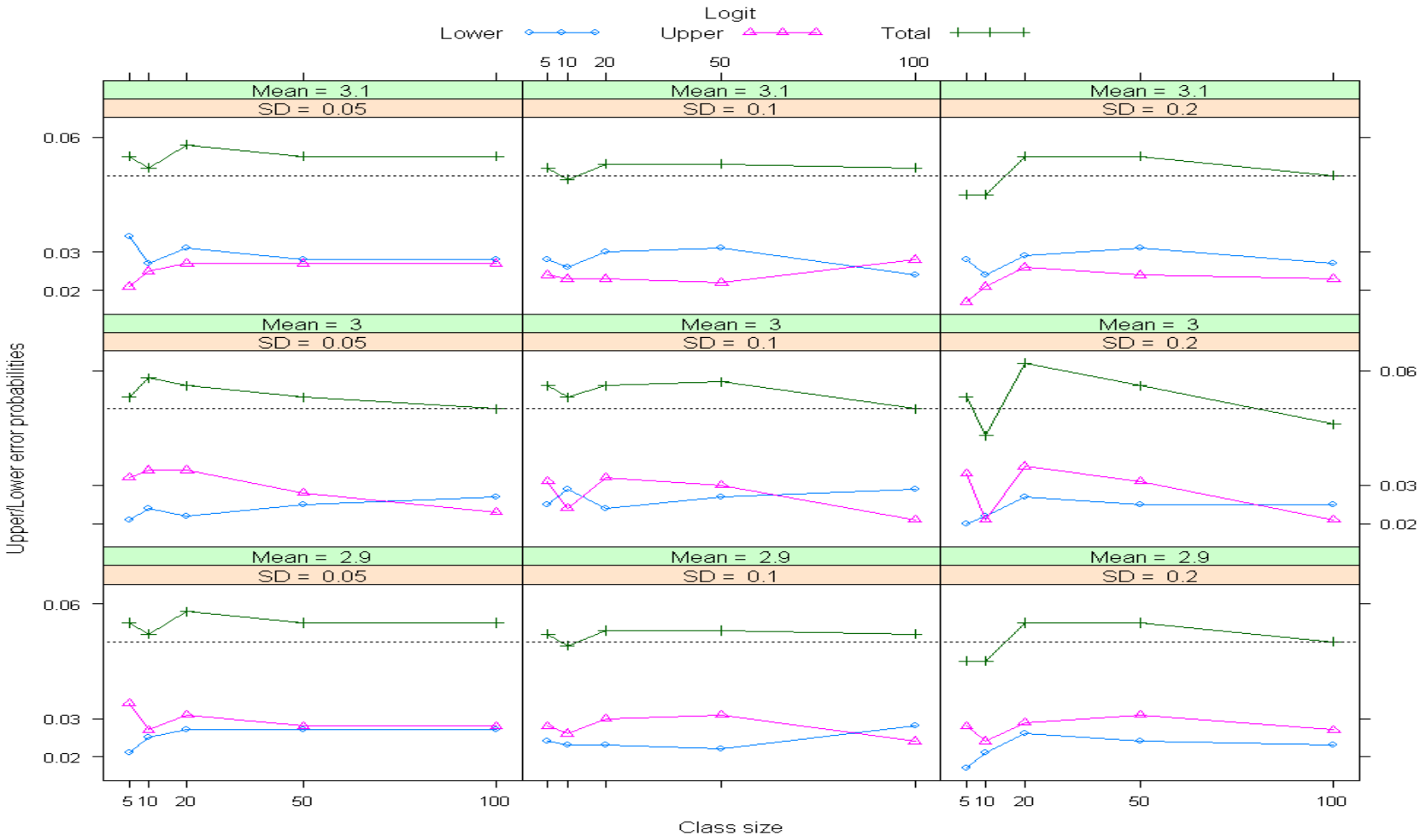

Figure 13. Logit upper/lower probabilities of the 95% CIs for the symmetrical distributions

Coverage error probabilities. The results for the coverage error probabilities are shown in Figure 15. These results are associated with the results obtained from the coverage probabilities criterion. The logit transformation method has almost zero coverage error for different sample sizes. We observe that the coverage error probabilities for the other methods are relatively large for small sample sizes. However, as the sample size increases, the estimated coverage error probability of these methods decreases.

Average width. The results for the average interval width are shown in Figure 16. We observe that the interval width of all methods are related to both the sample size, and the magnitude of the variation in the underlying distributions. As the variance of the sample increases, the estimated interval width increases and as the sample size decreases, the estimated interval width increases. All methods produced the same interval width with a sample size of 20 or more. However, with small variance, all methods obtain small estimates of interval length. In general, we also observe that the resulting interval of the logit method was marginally wider on average than the other methods when the sample size is 10 or less.

Upper and lower error probabilities. The results for the upper and lower error probabilities when sampling from a right-skewed distribution for the logit method are shown in Figure 17. The results show that the estimated confidence intervals of the logit method are asymmetric for different sample sizes in which the upper limit coverage errors are higher than the lower coverage error. However, all other methods (except Z method) obtain asymmetrical confidence interval with high skewed distribution and as the variance increases (Figures E2, E5, E8, E11) in Appendix E.

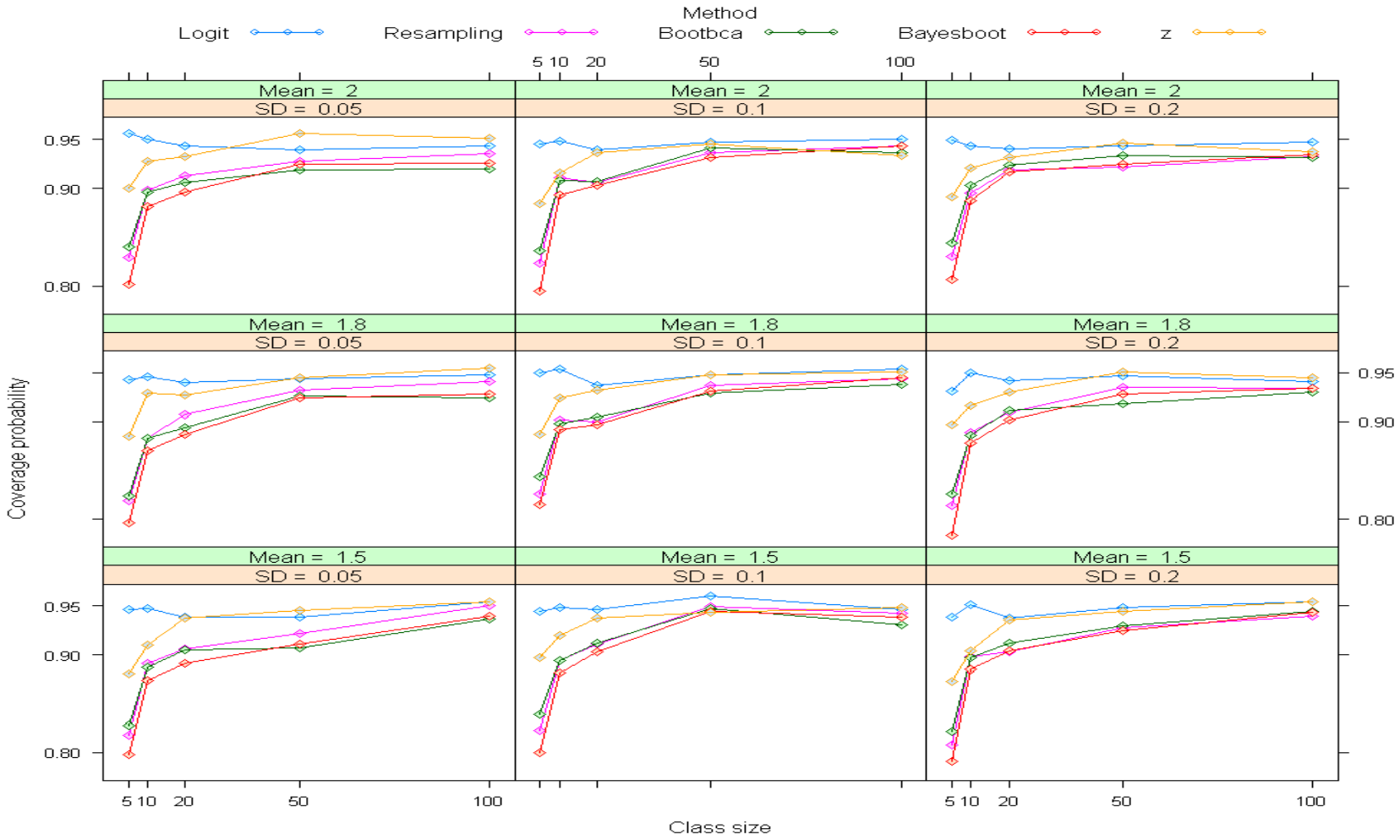

Figure 14. Coverage probabilities of the 95% CIs for the right-skewed distributions.

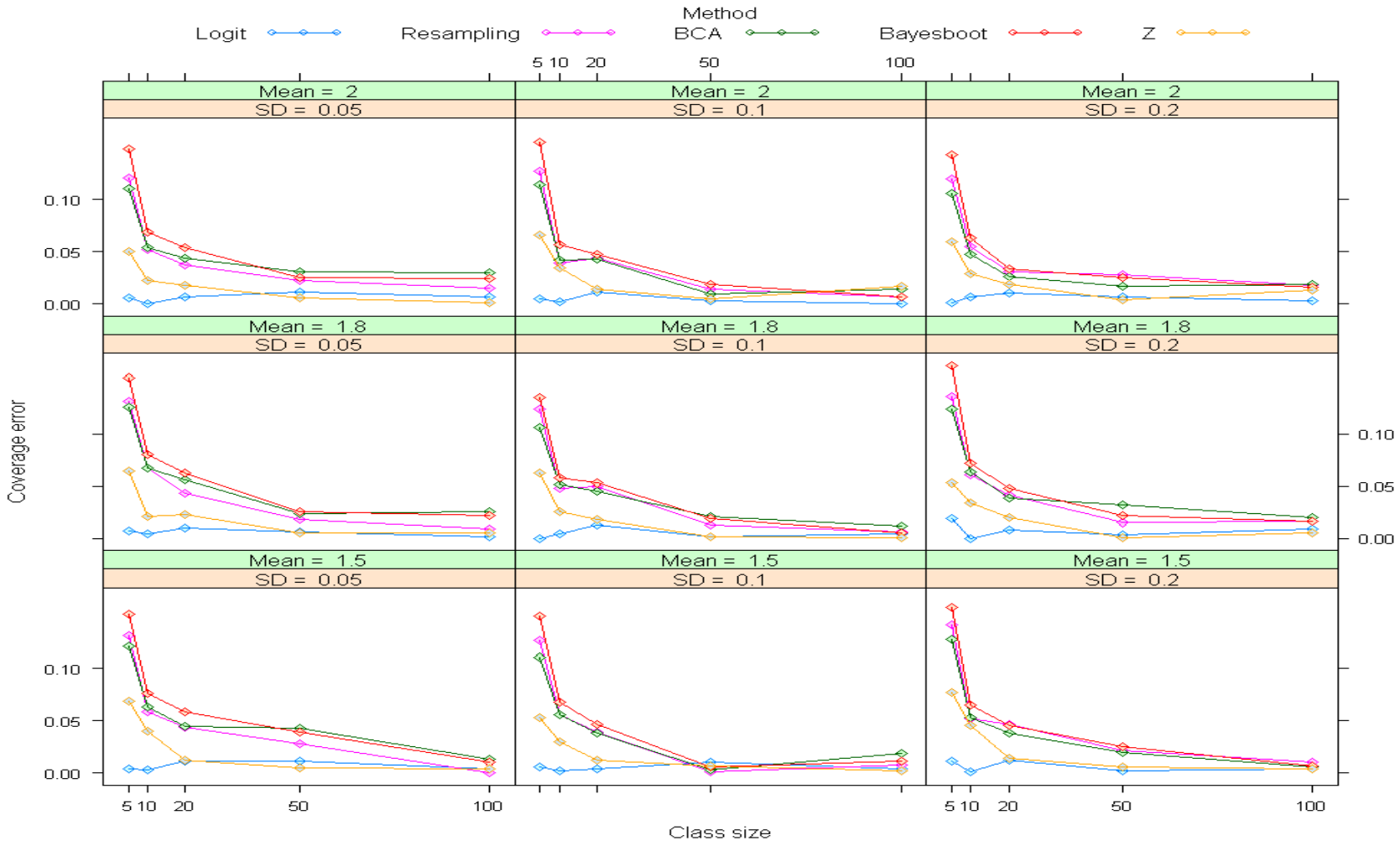

Figure 15. Coverage error probabilities of the 95% CIs for the right-skewed distributions.

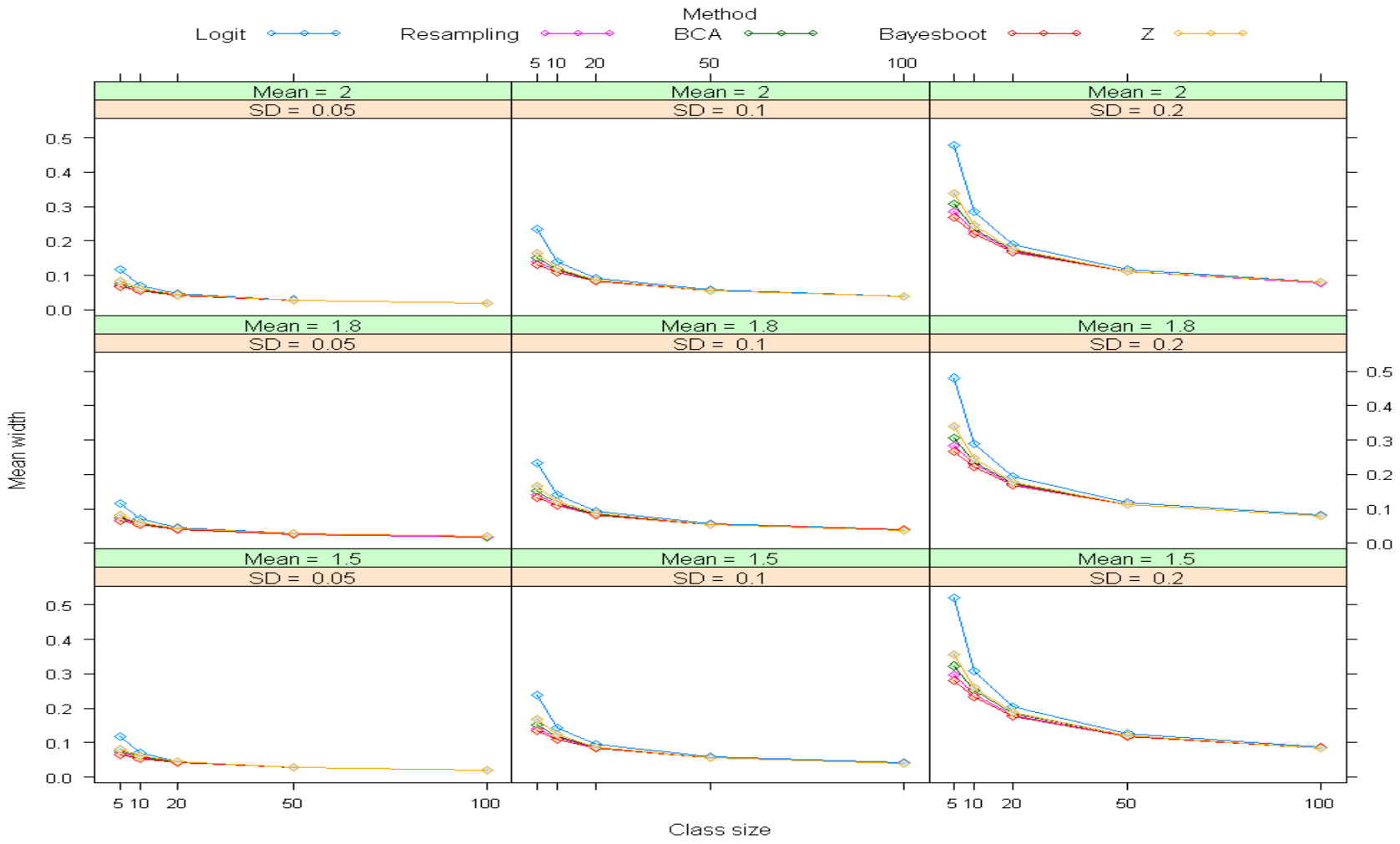

Figure 16. Average width of the 95% CIs for the right-skewed distributions.

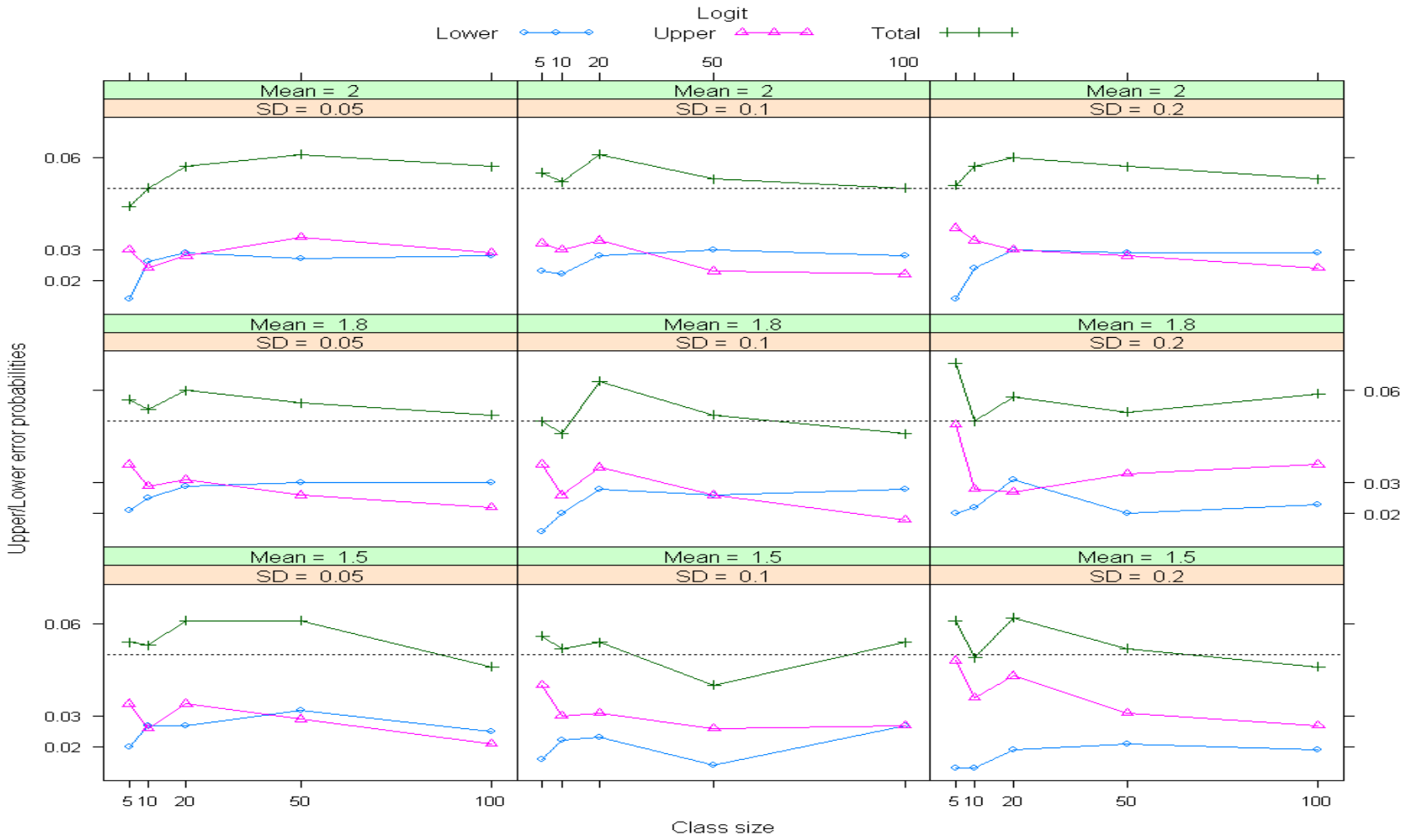

Figure 17. Logit upper/lower probabilities of the 95% CIs for the right-skewed distributions

Left-skewed distributions. In this section overall coverage probabilities, coverage errors, average widths, and upper and lower error probabilities of the resulting confidence intervals estimated by each method considered are given for the left-skewed distributions. The results are presented in plot form and reported in tables (see Tables D19-D27) in Appendix D.

Coverage probabilities. The results in Figure 18 suggest that when sampling from a left-skewed distribution, the logit transformation method has coverage probability close to the nominal 95% level and remains the same for different sample sizes. However, for small sample sizes, the estimated coverage probability of all methods (except logit transformation) are below the nominal 95% level. We observe that as the sample size increases, the estimated coverage probability of these methods increases. However, it is obvious that the Z distribution method outperformed the other bootstrap methods with sample size of 20 or more, because the coverage probabilities with this method tended to be higher than those with the other bootstrap methods. The resampling, BCa, and bayesboot need a sample size close to 100 or more to attain a value close to the nominal level 95%.

Coverage error probabilities. The results for the coverage error probabilities are shown in Figure 19. These results are associated with the results obtained from the coverage probabilities criterion. The logit transformation method has almost zero coverage error for different sample sizes. We observe that the coverage error probabilities for the other methods are relatively large for small sample size. However, as the sample size increases, the estimated coverage error probability of these methods decreases.

Average width. The results for the average width are shown in Figure 20. We observe that the interval width of all methods are related to both the sample size, and the magnitude of the variation in the underlying distributions. As the variance of the sample increases, the estimated interval width increases and as the sample size decreases, the estimated interval width increases. All methods produced the same interval width with sample size of 20 or more. However, with small variance, all methods obtain small estimates of interval length. In general, we also observe that the resulting interval of the logit method was marginally wider on average than the other methods when the sample size is 10 or less.

Upper and lower error probabilities. The results for the upper and lower error probabilities when sampling from a left-skewed distribution for logit method are shown in Figure 21. The results show that the estimated confidence intervals of the logit method are asymmetric for different sample sizes in which the lower limit coverage errors are higher than the upper coverage error. However, all other methods (except Z method) obtain asymmetrical confidence interval with high skewed distribution and as the variance increases (Figures E3, E6, E9, E12) in Appendix E.

The feasibility of the estimated procedures. Table 8 presents a comparison between the four interval estimation methods used in this study. All methods can be easily implemented and automated to any SET data. However, the bootstrap methods in general require extra programming effort, more computation time, and high computer resources (e.g., CPU and RAM). The logit transformation method requires less programming effort, less computation time, and less computer resources.

Table 8

The Feasibility of the Estimated Procedures

Criterion	Logit transformation	Resampling	BCA	Bayesboot	Z
Integration	Easy	Easy	Easy	Easy	Easy
Programming	Normal	Extra	Extra	Extra	Normal
Average Time	21 Seconds	47 Minutes	27 Minutes	19 Hours	25 Seconds

Note. The average time is calculated based on the simulation study.

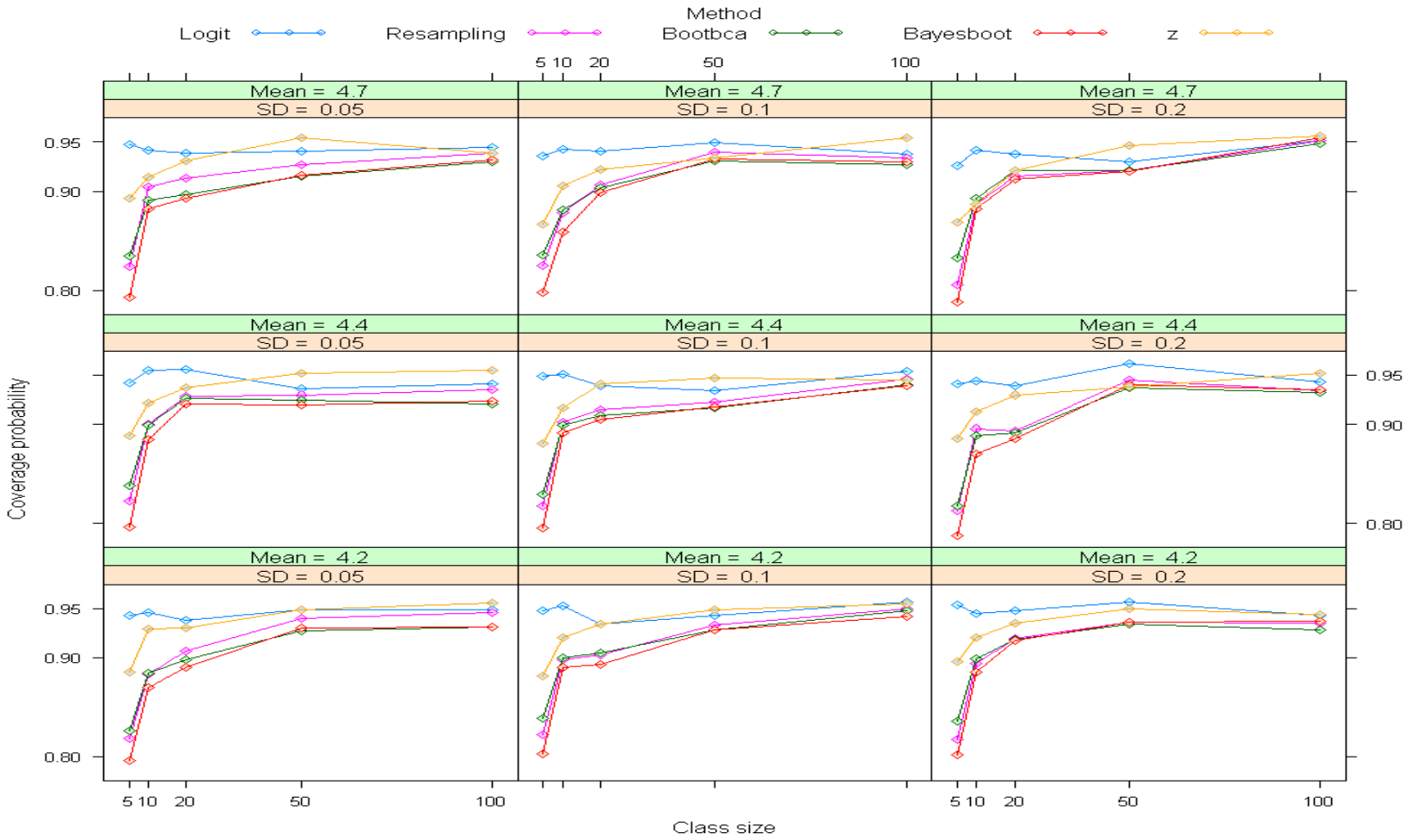

Figure 18. Coverage probabilities of the 95% CIs for the left-skewed distributions.

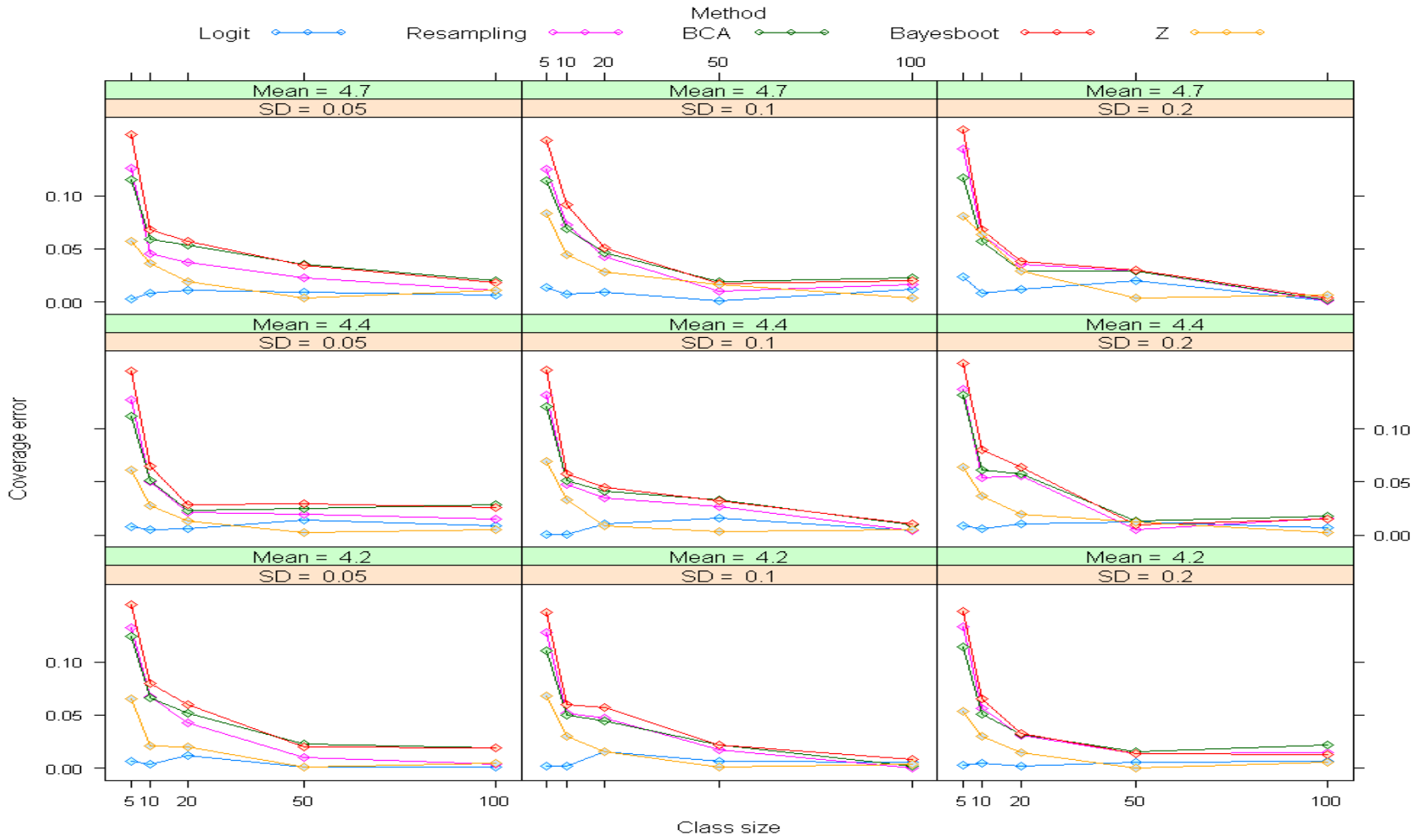

Figure 19. Coverage error probabilities of the 95% CIs for the left-skewed distributions.

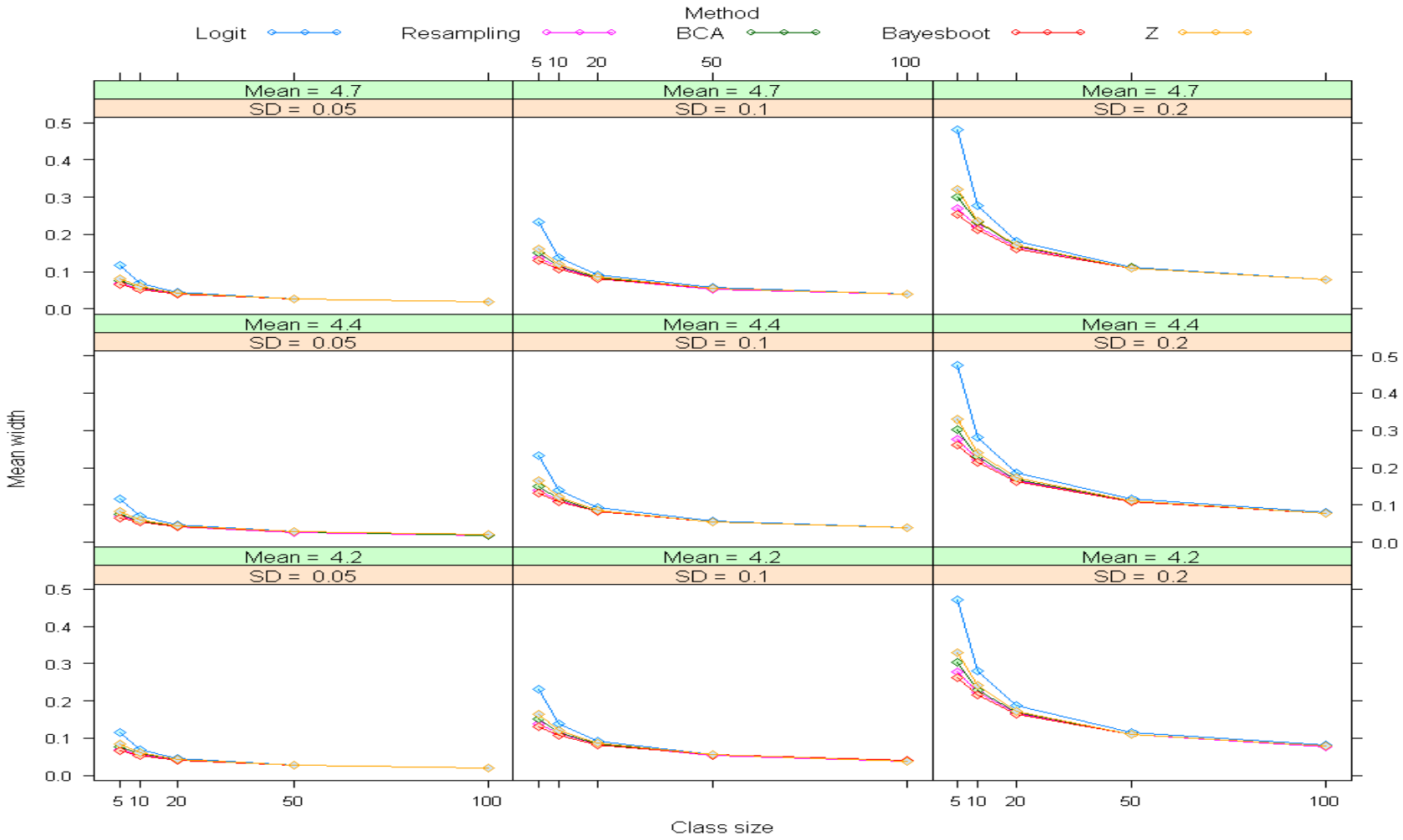

Figure 20. Average width of the 95% CIs for the left-skewed distributions.

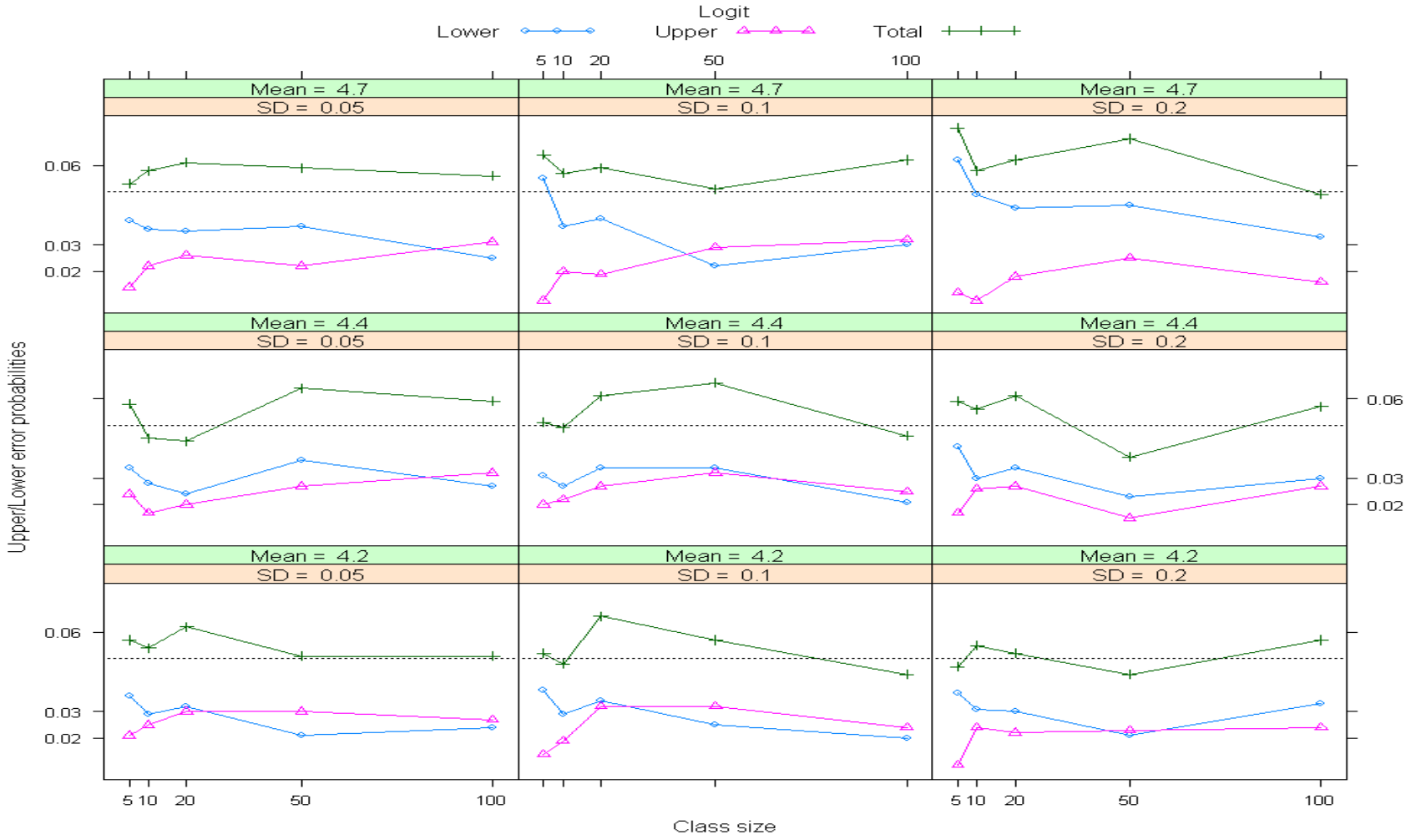

Figure 21. Logit upper/lower probabilities of the 95% CIs for the left-skewed distributions.

CHAPTER 5: Discussion

Reliability of the Estimated Class Means

Estimating the reliability of class means is an important initial step to justify and assess aggregated student ratings at the class level. Generalizability theory is preferred method over classical test theory for this purpose since it partitions the error variance into two or more components representing different likely sources of measurement error. In this study, we applied generalizability theory to estimate the reliability of class means and to project how the estimated reliability coefficients would likely vary if the number of items and responding students were increased or decreased.

The results show that the reliability estimates varied depending on whether the items were classified as a random or a fixed facet. More reliable results were obtained by treating items as fixed (e.g., .895). However, previous studies suggest that the decision to treat facets as random or fixed can only be made in the context of a particular study (Brennan, 2011; Kane & Brennan, 1977; Schweig, 2013). When items are classified as fixed, to obtain a reliability of .80 or higher the university needs at least 3 items and 15 respondents. When items are classified as random, to obtain a reliability of .80 or higher the university needs at least 3 items and 27 respondents.

The estimated reliability for the class-level mean using ICC_2 in this study supported the findings from the generalizability theory (ICC_2 = .867). As suggested by previous studies this estimate is desirable and indicating a sufficient degree of reliability of the class-mean ratings (Lüdtke et al., 2009; Nelson & Christ, 2016).

Uncertainty in Estimation of the Universe Mean for a Class

This dissertation used a simulation study to evaluate and compare four alternative methods of constructing confidence intervals for SET class means under different distributional

context. The four methods included (a) the logit transformation, (b) resampling, (c) BCA, and (d) Bayesboot. The classical Z method was included for comparison purposes. The criteria used to evaluate the five methods included (a) high coverage probability, (b) coverage error, (c) a narrow width, and (d) symmetric or asymmetric confidence intervals. However, selection of an appropriate confidence interval method does not need to satisfy all of these criteria simultaneously because it is difficult to find a confidence interval which satisfies all of these criteria at the same time. Thus, the researcher must decide which criterion is most important to the study and to pay the price for such a trade-off.

Findings from simulation study. The current simulation study led to six important findings regarding the five different methods used. These finding include the following:

1. When the data were symmetrically distributed, the five methods performed differently across all criteria. As Tables D1-D3 in Appendix D illustrated, all confidence intervals had coverage probabilities that were lower than the expected nominal 95% except for the logit transformation. The average interval widths were also different across the five methods when the sample size is less than 20 (Tables D4-D6 in Appendix D). All confidence intervals had a narrower width for small samples than the logit transformation interval. When the sample size is more than 20, all interval widths were fairly similar. It is expected that when the data are symmetrically distributed, the intervals would be symmetric and centered around the mean. The logit transformation interval was symmetrical for different sample sizes. All other methods require large sample size to obtain symmetrical confidence intervals (Tables D7-D9 in Appendix D).
2. When the data were nonormally distributed (either left or right skewed), the logit transformation method outperformed the other methods in coverage probability. As

Tables D10-D12 and Tables D19-D21 in Appendix D illustrated, all confidence intervals had coverage probabilities that were far lower than the expected nominal 95% except for the logit transformation and Z method with large sample sizes. The interval widths were also different across the five methods when the sample size was less than 20 (Tables D13-D15 and Tables D22-D24 in Appendix D). All confidence intervals had a narrower width for small samples than the logit transformation interval. When the sample size is more than 20, all interval widths were fairly similar. It is expected that when the data are nonnormally distributed, asymmetrical intervals would be estimated about the mean. Logit transformation intervals outperform all other methods to obtain asymmetrical interval for different sample sizes. All other methods obtain asymmetrical interval with high skewed distribution and as the variance increases (Tables D16-D18 and Tables D25-D27 in Appendix D).

3. In general, the coverage probabilities for confidence intervals became better as the sample size increases but still below the nominal 95% except for the logit transformation and Z method. The interval widths are related to both the sample size and the magnitude of the variation in the underlying distributions. As the variance of the sample increases and the sample size decreases, the estimated interval width becomes wider.
4. The findings indicate that the classical Z method performs better in terms of higher coverage probability and average interval width as the sample size increases. However, this method fails to produce an asymmetric interval about the estimated mean when the data are not normally distributed.
5. The findings indicate that the logit transformation method outperforms all other methods in terms of higher coverage probability and symmetric/asymmetric interval about the

mean, but the average widths are much wider than the average width intervals produced by the other methods when the sample size is small (e.g., 10 or less).

6. The logit transformation method is preferable to other methods in term of the feasibility of the estimation procedures because it requires less time and computer resources to estimate a confidence interval for each class.

General Discussion

Student evaluations of teaching effectiveness (SET) in higher education are commonly used in making important decisions regarding faculty promotion, raises, and tenure. Professors and administrators who use SET for such decisions have a need to be assured that the resulting ratings are a reliable indicator of a teacher's effectiveness. Generalizability theory is suggested to justify and assess aggregated student ratings at the class level (Lüdtke et al., 2006; Nelson & Christ, 2016). Consistent with previous studies, the results of our generalizability theory analysis indicated that the effect of increasing the number of responding students' has a greater effect than increasing the number of items in the SET instrument (Gilmore et al., 1978; Ibrahim, 2011). This finding suggests that the ratings by a small number of respondents in classes are most likely to be too unreliable to be used as a basis for making important personnel decisions.

Few published SET studies have paid attention to developing concepts and procedures to summarize information about the lack of precision associated with using the observed class mean as an estimate of the corresponding universe mean. Boysen (2015a) declared that "means are only an estimate of true scores; thus, teaching evaluation means should be interpreted as estimates falling within possible range of scores rather than a representation of true teaching competency" (p.151). The purpose for reporting a confidence interval along with the estimated class means is to emphasize that the reported mean is a fallible estimate of the unknown true

mean for a given class and to describe how accurate or inaccurate that estimate is. Hence, reporting the confidence interval is intended to help the users of SET results from over-interpreting small differences in the observed class means as being indicative of true differences in teachers. When the confidence intervals for two or more teachers overlap, any differences in their observed class means may be due to measurement error or sampling error. Hence, we cannot be confident in inferring that their true scores are actually different.

Findings from the current study suggest that the logit transformation method outperforms other methods in coverage probabilities and in terms of the degree and direction of asymmetry of the interval about the mean with small class sizes and nonnormal distributions. This finding is consistent with the results obtained by Choi et al. (2013) regarding of the performance of a simple Wald-based CIs using transformation. The resampling method showed poor coverage probability with small sample sizes and nonnormal distributions. The results from the current simulation are consistent with Ghosh and Polansky (2014) in terms of both coverage probability and interval width. The results of BCA method results are consistent with previous studies in showing low coverage probabilities but a narrow interval width with small sample sizes and skewed distribution (Banik & Kibria, 2010; Puth et. al, 2015; Wang, 2001). Likewise, the Bayesboot method provides similar results to non-parametric bootstrap methods with small samples sizes and skewed distributions. This consistent with previous research by O'Hagan and Stevens (2003). The classical Z method provides similar results to logit transformation method and outperforms the other bootstrap methods. However, a major limitation of this method is the failing to estimate asymmetry interval when the data were nonormally distributed. This limitation provides estimates that may exceed the bounded range of the distribution.

Limitations

The primary limitations of this study are the small size of some classes and the low response rate within many classes. These limitations affect the accuracy of margin of error as well as the reliability of the class means in the case of small size classes and classes with low response rates.

Generalizability theory was used in this study to estimate different variance components. However, there might be other hidden facets that were not considered such as the subject area or instructor. These potential sources of variability were not included because of the unavailability of data. Accounting for more variance errors would potentially impact the results and the interpretation of the data.

Using uninformative prior with Bayesboot method also introduces another limitation. The uninformative prior produces almost the same estimates compare to other bootstrap methods used in this study.

Another limitation is related to the computer resources and the time needed for utilizing some methods (e.g., Bayesboot) in the analysis, taking into account the huge number of classes that need to included in such analysis.

Recommendations and Future Research

The following recommendations are offered based on the results of this study and the accompanying review of related published literature on SET ratings. Institutions which use SET ratings should implement the following actions:

1. Routinely incorporate uncertainty information through the use of appropriate confidence intervals whenever SET ratings are presented and displayed.

2. Use the logit transformation procedure to compute confidence intervals for the mean SET ratings for each class.
3. Because of the lower reliability of the estimated means obtained from classes with less than 15 respondents, employ extra caution when interpreting the mean ratings for these classes and treat the resulting estimates as tentative unless additional information is obtained.
4. Consider ways to increase the response rate in all classes, especially in classes with enrollments of 20 or less.
5. Conduct research to examine the potential effects of bias due to nonresponse in estimating the mean SET ratings of classes.
6. Consider the appropriateness of combining SET ratings across multiple semesters for a given class.
7. Use multilevel and generalizability analyses to gauge the appropriateness of aggregating results at other organizational levels (e.g., programs, departments, colleges), including the estimation of standard errors and confidence intervals at these other levels of aggregation.

References

AbouRizk, S. M., Halpin, D. W., & Wilson, J. R. (1991). Visual interactive fitting of beta distributions. *Journal of Construction Engineering and Management, 117*(4), 589-605.

Abrami, P. C. (2001). Improving judgments about teaching effectiveness using teacher rating forms. *New Directions for Institutional Research*, *109*, 59–87.

Abu-Shawiesh, M. O. A., Banik, S., & Kibria, B. G. (2011). A simulation study on some confidence intervals for the population standard deviation. *SORT-Statistics and Operations Research Transactions, 35*(2), 83-102.

Agresti, A., & Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician, 52*(2), 119-126.

Alfaro, M. E., Zoller, S., & Lutzoni, F. (2003). Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. *Molecular Biology and Evolution, 20*(2), 255-266.

Alkharusi, H. A. (2012). A generalizability approach to the measurement of score reliability of the Teacher Assessment Literacy Questionnaire. *Journal of Studies in Education, 2*(2), 157-164.

Baath, R. (2016). *Bayesboot: An implementation of Rubin's (1981) Bayesian bootstrap*. R package.

Banik, S., & Kibria, B. G. (2010). Comparison of some parametric and nonparametric type one sample confidence intervals for estimating the mean of a positively skewed distribution. *Communications in Statistics—Simulation and Computation®, 39*(2), 361-389.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238-246.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*(3), 588-606.

Beran, T., & Violato, C. (2005). Ratings of university teacher instruction: How much do student and course characteristics really matter? *Assessment & Evaluation in Higher Education, 30*(6), 593-601. doi:10.1080/02602930500260688

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein & S. W. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 349 –381). San Francisco, CA: Jossey-Bass.

Boysen, G. A. (2015a). Significant interpretation of small mean differences in student evaluations of teaching despite explicit warning to avoid overinterpretation. *Scholarship of Teaching and Learning in Psychology, 1*(2), 150-162.

Boysen, G. A. (2015b). Uses and misuses of student evaluations of teaching
the interpretation of differences in teaching evaluation means irrespective of statistical information. *Teaching of Psychology, 42*(2), 109-118. doi:10.1177/0098628315569922

Boysen, G. A. (2016). Using student evaluations to improve teaching: Evidence-based recommendations. *Scholarship of Teaching and Learning in Psychology, 2*(4), 273-284.

Boysen, G. A., Kelly, T. J., Raesly, H. N., & Casner, R. W. (2014). The (mis) interpretation of teaching evaluations by college faculty and administrators. *Assessment & Evaluation in Higher Education, 39*(6), 641-656.

Brennan, R. L. (1992). Generalizability theory. *Educational Measurement: Issues and Practice, 11*(4), 27-34

Brennan, R. L. (2001). *Generalizability theory: Statistics for social science and public policy*. New York, NY: Springer-Verlag.

Brennan, R. L. (2011). Generalizability theory and classical test theory. *Applied Measurement in Education, 24*(1), 1-21.

Briesch, A. M., Swaminathan, H., Welsh, M., & Chafouleas, S. M. (2014). Generalizability theory: A practical guide to study design, implementation, and interpretation. *Journal of School Psychology, 52*(1), 13-35.

Briggs, A. H., Wonderling, D. E., & Mooney, C. Z. (1997). Pulling cost-effectiveness analysis up by its bootstraps: A non-parametric approach to confidence interval estimation. *Health Economics, 6*(4), 327-340.

Brondino, M., Pasini, M., & da Silva, S. C. A. (2013). Development and validation of an integrated organizational safety climate questionnaire with multilevel confirmatory factor analysis. *Quality & Quantity, 47*(4), 2191-2223.

Brown, T. A. (2014). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Publications.

Calzada, M. E., & Gardner, H. (2011). Confidence intervals for the mean: To bootstrap or not to bootstrap. *Mathematics and Computer Education, 45*(1), 28-38.

Campbell, J. P. (2005). *Evaluating teacher performance in higher education: The value of student ratings* (Doctoral dissertation, University of Central Florida Orlando, Florida).

Carpenter, J., & Bithell, J. (2000). Bootstrap confidence intervals: When, which, what? A practical guide for medical statisticians. *Statistics in Medicine, 19*(9), 1141-1164.

Chen, X. (2008). A simple formula for constructing confidence interval for the mean of bounded random variables. *arXiv Preprint arXiv:0802.3458,*

Chihara, L. & Hesterberg, T. (2011*). Mathematical statistics with resampling and R.* Hoboken, NJ: Wiley.

Choi, B. Y., Fine, J. P., & Brookhart, M. A. (2013). Practicable confidence intervals for current status data. *Statistics in Medicine, 32*(8), 1419-1428.

Cooley, D. (2013). Return periods and return levels under climate change. In Extremes in a Changing Climate (pp. 97-114). Netherlands: Springer.

Cordeiro, G. M., & de Castro, M. (2011). A new family of generalized distributions. *Journal of Statistical Computation and Simulation, 81*(7), 883-898.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *Theory of generalizability for scores and profiles. the dependability of behavioral measurements.* New York, NY: Wiley.

Cumming, G., & Fidler, F. (2009). Confidence intervals: Better answers to better questions. *Zeitschrift für Psychologie/Journal of Psychology, 217*(1), 15-26.

Dedrick, R. F., & Greenbaum, P. E. (2011). Multilevel confirmatory factor analysis of a scale measuring interagency collaboration of children's mental health agencies. *Journal of Emotional and Behavioral Disorders, 19*(1), 27-40.

DeFrain, E. (2016). *An analysis of differences in non-instructional factors affecting teacher-course evaluations over time and aCPoss disciplines* (Doctoral dissertation, University of Arizona, Arizona)

DiCiccio, T. J., & Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science, 11*(3), 189-212.

Dixon, P. M. (2002) Bootstrap resampling. *Encyclopedia of Environmetrics, 1*, 212-220.

Dodeen, H. (2013). Validity, reliability, and potential bias of short forms of students' evaluation of teaching: The case of UAE university. *Educational Assessment, 18*(4), 235-250.

Dyer, N. G., Hanges, P. J., & Hall, R. J. (2005). Applying multilevel confirmatory factor analysis techniques to the study of leadership. *The Leadership Quarterly, 16*(1), 149-167.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics.7*: 1-26.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association, 82*(397), 171-185.

Efron, B. (2015). Frequentist accuracy of Bayesian estimates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 77*(3), 617-646.

Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science, 1*(1), 54-75.

Farnum, N. R., & Stanton, L. W. (1987). Some results concerning the estimation of beta distribution parameters in PERT. *Journal of the Operational Research Society, 38*(3), 287-290.

Fernandez, J., Mateo, M. A., & Muniz, J. (1998). Is there a relationship between class size and student ratings of teaching quality? *Educational and Psychological Measurement, 58*(4), 596-604.

Franklin, J. (2001). Interpreting the numbers: Using a narrative to help others read student evaluations of your teaching accurately. *New Directions for Teaching and Learning, 2001*(87), 85-100.

Frenzel, A. C., Goetz, T., Lüdtke, O., Pekrun, R., & Sutton, R. E. (2009). Emotional transmission in the classroom: Exploring the relationship between teacher and student enjoyment. *Journal of Educational Psychology, 101*(3), 705-716.

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods, 19*(1), 72-91.

Ghosh, S., & Polansky, A. M. (2014). Smoothed and iterated bootstrap confidence regions for parameter vectors. *Journal of Multivariate Analysis, 132*, 171-182.

Ghosh, S., & Polansky, A. M. (2016). New bootstrap confidence intervals for means of positively skewed distributions. *Communications in Statistics-Theory and Methods, 45*(23), 6915-6927.

Gillmore, G. M., Kane, M. T., & Naccarato, R. W. (1978). The generalizability of student ratings of instruction: Estimation of the teacher and course components. *Journal of Educational Measurement*, *15*(1), 1-13.

Gross, J., Lakey, B., Edinger, K., Orehek, E., & Heffron, D. (2009). Person perception in the college classroom: Accounting for taste in students' evaluations of teaching effectiveness. *Journal of Applied Social Psychology, 39*(7), 1609-1638.

Gu, J., Ghosal, S., & Roy, A. (2008). Bayesian bootstrap estimation of ROC curve. *Statistics in Medicine, 27*(26), 5407-5420.

Hair, J. F., Jr., Anderson, R. E., Tatham, R. L., & Black, W. C. (1995). *Multivariate data analysis* (3rd ed.). New York, NY: Macmillan Publishing Company.

Hobson, S. M., & Talbot, D. M. (2001). Understanding student evaluations: What all faculty should know. *College Teaching, 49*(1), 26-31.

Hu, C., Yeilding, N., Davis, H. M., & Zhou, H. (2011). Bounded outcome score modeling: Application to treating psoriasis with ustekinumab. *Journal of Pharmacokinetics and Pharmacodynamics, 38*(4), 497-517.

Hu, L., & Bentler, P. M. (1999). Cutoff CPiteria for fit indexes in covariance structure analysis: Conventional CPiteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1-55.

Huang, Z., Li, J., Cheng, C., Cheung, C., & Wong, T. (2016). Bayesian reclassification statistics for assessing improvements in diagnostic accuracy. *Statistics in Medicine, 35*(15), 2574–2592

Hurairah, A., Akma Ibrahim, N., Bin Daud, I., & Haron, K. (2006). Approximate confidence interval for the new extreme value distribution. *Engineering Computations, 23*(2), 139-153.

Ibrahim, A. M. (2011). Using generalizability theory to estimate the relative effect of class size and number of items on the dependability of student ratings of instruction. *Psychological Reports, 109*(1), 252-258.

James, D. E., Schraw, G., & Kuch, F. (2015). Using the sampling margin of error to assess the interpretative validity of student evaluations of teaching. *Assessment & Evaluation in Higher Education, 40*(8), 1123-1141.

Jeon, M., Lee, G., Hwang, J., & Kang, S. (2009). Estimating reliability of school-level scores using multilevel and generalizability theory models. *Asia Pacific Education Review, 10*(2), 149-158.

Kane, M. T., & Brennan, R. L. (1977). The generalizability of class means. *Review of Educational Research, 47*(2), 267-292.

Kane, M. T., Gillmore, G. M., & CPooks, T. J. (1976). Student evaluations of teaching: The generalizability of class means. *Journal of Educational Measurement, 13*(3), 171-183.

Karian, Z. A., & Dudewicz, E. J. (2011). Fitting data and distributions with the GLD and kappa distributions using percentiles and L-moments. *Journal of Combinatorics, Information & System Sciences, 36*(1-4), 285-324.

Kelley, K. (2005). The effects of nonnormal distributions on confidence intervals around the standardized mean difference: Bootstrap and parametric confidence intervals. *Educational and Psychological Measurement, 65*(1), 51-69.

Kelley, K., & Rausch, J. R. (2006). Sample size planning for the standardized mean difference: accuracy in parameter estimation via narrow confidence intervals. *Psychological Methods, 11*(4), 363-385.

Kong, J., Parker, J., & Sul, D. (2014). *Improved Two-Sample Comparisons for Bounded Data,* (Unpublished doctoral dissertation), University of Texas at Dallas, Dallas, TX.

Kulik, J. A. (2001). Student ratings: Validity, utility, and controversy. *New Directions for Institutional Research, 2001*(109), 9-25.

Lang, J. W., & Kersting, M. (2007). Regular feedback from student ratings of instruction: Do college teachers improve their ratings in the long run? *Instructional Science, 35*(3), 187-205.

Lazar, R., Meeden, G., & Nelson, D. (2008). A noninformative Bayesian approach to finite population sampling using auxiliary variables. *Survey Methodology, 34*(1), 51-64.

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11(4)*, *815-852*

Lesaffre, E., Rizopoulos, D., & Tsonaka, R. (2007). The logistic transform for bounded outcome scores. *Biostatistics, 8*(1), 72-85.

Liu, O. (2012). Student evaluation of instruction: In the new paradigm of distance education. *Research in Higher Education, 53*(4), 471-486.

Liu, X. S. (2009). Sample size and the width of the confidence interval for mean difference. *British Journal of Mathematical and Statistical Psychology, 62*(2), 201-215.

Lo, A. Y. (1988). A Bayesian bootstrap for a finite population. *The Annals of Statistics,16*(4), 1684-1695.

Lüdtke, O., Robitzsch, A., Trautwein, U., & Kunter, M. (2009). Assessing the impact of learning environments: How to use student ratings of classroom or school characteristics in multilevel modeling. *Contemporary Educational Psychology, 34*(2), 120-131.

Lüdtke, O., Trautwein, U., Kunter, M., & Baumert, J. (2006). Reliability and agreement of student ratings of the classroom environment: A reanalysis of TIMSS data. *Learning Environments Research, 9*(3), 215-230.

Luh, W., & Guo, J. (2001). Transformation works for non-normality? on one-sample transformation trimmed t methods. *British Journal of Mathematical and Statistical Psychology, 54*(2), 227-236.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*(2), 130-149.

Mammen, E. (1992). *When does bootstrap work? Asymptotic results and simulations Lecture Notes in Statistics*, 77. New York: Springer.

Marsh, H. W. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In R. P. Perry & J. C. Smart (Eds.), *The*

scholarship of teaching and learning in higher education (pp. 319–383). Netherlands: Springer.

Marsh, H. W., Ginns, P., Morin, A. J., Nagengast, B., & Martin, A. J. (2011). Use of student ratings to benchmark universities: Multilevel modeling of responses to the Australian Course Experience Questionnaire (CEQ). *Journal of Educational Psychology, 103*(3), 733-748.

Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J., Abduljabbar, A. S., & Koller, O. (2012). Classroom climate and contextual effects: Conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist, 47*(2), 106-124.

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist, 52*(11), 1187-1197.

Meeden, G. (1999). Interval estimators for the population mean for skewed distributions with a small sample size. *Journal of Applied Statistics, 26*(1), 81-96.

Miller, J. M., & Penfield, R. D. (2005). Using the score method to construct asymmetric confidence intervals: An SAS program for content validation in scale development. *Behavior Research Methods, 37*(3), 450-452.

Morin, A. J., Marsh, H. W., Nagengast, B., & Scalas, L. F. (2014). Doubly latent multilevel analyses of classroom climate: An illustration. *The Journal of Experimental Education, 82*(2), 143-167.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research, 22*(3), 376-398.

Muthén, L. K., & Muthén, B. O. (2015). *Mplus: statistical analysis with latent variables.* Los Angeles, CA: Muthén & Muthén.

Narayanan, A., Sawaya, W. J., & Johnson, M. D. (2014). Analysis of differences in nonteaching factors influencing student evaluation of teaching between engineering and business classrooms. *Decision Sciences Journal of Innovative Education, 12*(3), 233-265.

Nelson, C. H., & Preckel, P. V. (1989). The conditional beta distribution as a stochastic production function. *American Journal of Agricultural Economics, 71*(2), 370-378.

Nelson, P. M., & Christ, T. J. (2016). Reliability and agreement in student ratings of the class environment. *School Psychology Quarterly, 31(3)*, 419-430.

Newcombe, R. G. (1998). Two-sided confidence intervals for the single proportion: Comparison of seven methods. *Statistics in Medicine, 17*(8), 857-872.

Newman, D. A., & Sin, H. (2007). How do missing data bias estimates of within-group agreement? Sensitivity of SDWG, CVWG, rWG (J), rWG (J)*, and ICC to systematic nonresponse. *Organizational Research Methods, 12*(1), 113-147.

Nicholls, A. (2014). Confidence limits, error bars and method comparison in molecular modeling. Part 1: the calculation of confidence intervals. *Journal of Computer-Aided Molecular Design, 28*(9), 887-918.

Nixon, R. M., Wonderling, D., & Grieve, R. D. (2010). Non-parametric methods for cost-effectiveness analysis: The central limit theorem and the bootstrap compared. *Health Economics, 19*(3), 316-333.

Nulty, D. D. (2008). The adequacy of response rates to online and paper surveys: What can be done? *Assessment & Evaluation in Higher Education, 33*(3), 301-314.

O'Hagan, A., & Stevens, J. W. (2003). Assessing and comparing costs: How robust are the bootstrap and methods based on asymptotic normality? *Health Economics, 12*(1), 33-49.

Ory, J. C. (2001). Faculty thoughts and concerns about student ratings. *New directions for teaching and learning, 2001*(87), 3-15.

Ory, J. C., & Ryan, K. (2001). How do student ratings measure up to a new validity framework? . *New directions for institutional research, 2001*(109), 27-44.

Penfield, R. D. (2003). A score method of constructing asymmetric confidence intervals for the mean of a rating scale item. *Psychological Methods*, *8*(2), 149-163.

Penfield, R. D., & Miller, J. M. (2004). Improving content validation studies using an asymmetric confidence interval for the mean of expert ratings. *Applied Measurement in Education, 17*(4), 359-370.

Pincus, H. S., & Schmelkin, L. P. (2003). Faculty perceptions of academic dishonesty: A multidimensional scaling analysis. *Journal of Higher Education, 74*(2), 196-209.

Praetorius, A., Lenske, G., & Helmke, A. (2012). Observer ratings of instructional quality: Do they fulfill what they promise? *Learning and Instruction, 22*(6), 387-400.

Proctor, T. P. (2003). The Psychometrics of Student Evaluations of Instructors and Courses at Eastern Michigan University. (Unpublished Master's Thesis, Eastern Michigan University, Ypsilanti, Michigan)

Puth, M. T., Neuhäuser, M., & Ruxton, G. D. (2015). On the variety of methods for calculating confidence intervals by bootstrapping. *Journal of Animal Ecology, 84*(4), 892-897.

Qin, G., & Hotilovac, L. (2008). Comparison of non-parametric confidence intervals for the area under the ROC curve of a continuous-scale diagnostic test. *Statistical Methods in Medical Research, 17(2)*, 207-221.

R Core Team (2016). R: *A language and environment for statistical computing*. R Foundation For Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Radmacher, S. A., & Martin, D. J. (2001). Identifying significant predictors of student evaluations of faculty through hierarchical regression analysis. *Journal of Psychology, 135*(3), 259-268.

Raykov, T., & Marcoulides, G. A. (2006). On multilevel model reliability estimation from the perspective of structural equation modeling. *Structural Equation Modeling, 13*(1), 130-141.

Rantanen, P. (2013). The number of feedbacks needed for reliable evaluation. A multilevel analysis of the reliability, stability and generalisability of students' evaluation of teaching. *Assessment & Evaluation in Higher Education, 38*(2), 224-239.

Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. Thousand Oaks, CA: Sage.

Rubin, D. B. (1981). The Bayesian bootstrap. *The Annals of Statistics, 9*(1), 130-134.

Sappakitkamjorn, J., & Niwitpong, S. (2013). Confidence intervals for the coefficients of variation with bounded parameters. *International Journal of Mathematical, Computational, Physical, Electrical and Computer Engineering, 7*(9), 198-203.

Savoy, J. (1997). Statistical inference in retrieval effectiveness evaluation. *Information Processing & Management, 33*(4), 495-512.

Schweig, J. (2013). *Measurement error in multilevel models of school and classroom environments: Implications for reliability, precision, and prediction*. (CRESST Report 828.) Los Angeles: University of California at Los Angeles, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Schweig, J. D. (2016). Moving beyond means: Revealing features of the learning environment by investigating the consensus among student ratings. *Learning Environments Research, 19*(3), 441-462.

Shi, W., & Golam Kibria, B. M. (2007). On some confidence intervals for estimating the mean of a skewed population. *International Journal of Mathematical Education in Science and Technology, 38*(3), 412-421.

Sim, J., & Reid, N. (1999). Statistical inference by confidence intervals: Issues of interpretation and utilization. *Physical Therapy, 79*(2), 186-195.

Sojka, J., Gupta, A. K., & Deeter-Schmelz, D. (2002). Student and faculty perceptions of student evaluations of teaching. *College Teaching, 50*(2), 44-49.

Swift, M. B. (2009). Comparison of confidence intervals for a Poisson mean–further considerations. *Communications in Statistics—Theory and Methods, 38*(5), 748-759.

Tibshirani R, Leisch F. 2015. bootstrap: Functions for the Book "An Introduction to the Bootstrap". R package version 2015.2. Available from: http://CRAN.R-project.org/package=bootstrap

Ting, K. (2000). A multilevel perspective on student ratings of instruction: Lessons from the Chinese experience. *Research in Higher Education, 41*(5), 637-661.

Tong, L., Chang, C., Jin, S., & Saminathan, R. (2012). Quantifying uncertainty of emission estimates in national greenhouse gas inventories using bootstrap confidence intervals. *Atmospheric Environment, 56*, 80-87.

Van Albada, S. J., & Robinson, P. A. (2007). Transformation of arbitrary distributions to the normal distribution with application to EEG test–retest reliability. *Journal of Neuroscience Methods, 161*(2), 205-211.

Wang, F. K. (2001). Confidence interval for the mean of non-normal data. *Quality and Reliability Engineering International, 17*(4), 257-267.

Warton, D. I., & Hui, F. K. (2011). The arcsine is asinine: The analysis of proportions in ecology. *Ecology, 92*(1), 3-10.

Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2006). Reliability coefficients and generalizability theory. In C.R. Rao & S. Sinharay (Eds.) *Handbook of statistics* (pp. 81-124). Amsterdam: Elsevier.

Wei, X., & Haertel, E. (2011). The effect of ignoring classroom-level variance in estimating the generalizability of school mean scores. *Educational Measurement: Issues and Practice, 30*(1), 13-22.

Weng, C. (1989). On a second-order asymptotic property of the Bayesian bootstrap mean. *Annals of Statistics, 17*(2):705–710.

Wilcox, R. R. (2001). *Fundamentals of modern statistical methods: Substantially improving power and accuracy*. New York, NY: Springer.

Williamson, M., & Gaston, K. (1999). A simple transformation for sets of range sizes. *Ecography, 22*(6), 674-680.

Willink, R. (2007). On a confidence interval for the mean of an asymmetric distribution. *Communications in Statistics - Theory and Methods, 36*(12), 2235-2236.

Woehr, D. J., Loignon, A. C., Schmidt, P. B., Loughry, M. L., & Ohland, M. W. (2015). Justifying aggregation with consensus-based constructs: A review and examination of cutoff values for common aggregation indices. *Organizational Research Methods, 18*(4), 704-737.

Wu, J., Wong, A. C. M., & Jiang, G. (2003). Likelihood-based confidence intervals for a log-normal mean. *Statistics in Medicine, 22*(11), 1849-1860.

Zhou, X. H., & Gao, S. (1997). Confidence intervals for the log-normal mean. *Statistics in Medicine, 16*(7), 783-790.

Zou, G. Y., & Donner, A. (2008). Construction of confidence limits about effect measures: A general approach. *Statistics in Medicine, 27*(10), 1693-1702. doi:10.1002/sim.3095

Zumrawi, A. A., Bates, S. P., & Schroeder, M. (2014). What response rates are needed to make reliable inferences from student evaluations of teaching? *Educational Research and Evaluation, 20*(7-8), 557-563.

APPENDIX A:

Specimen Copy of the New Student Evaluation of Teaching Form Used at Brigham Young University

1. How effective was this instructor (not the TA) in helping students who indicated a need for assistance?
    a. Not at all effective
    b. Not very effective
    c. Moderately effective
    d. Effective
    e. Very effective

2. How effective was the instructor (not the TA) in providing meaningful opportunities and encouragement for you to actively participate in the learning process?
    a. Not at all effective
    b. Not very effective
    c. Moderately effective
    d. Effective
    e. Very effective

3. How effective was the instructor (not the TA) in teaching challenging concepts or skills?
    a. Not at all effective
    b. Not very effective
    c. Moderately effective
    d. Effective
    e. Very effective

4. How effective was this instructor in demonstrating respect for students and their opinions, questions, or concerns?
    a. Not at all effective
    b. Not very effective
    c. Moderately effective
    d. Effective
    e. Very effective

5. How effective was the instructor in organizing the course content to enhance learning?
    a. Not at all effective
    b. Not very effective
    c. Moderately effective
    d. Effective
    e. Very effective

APPENDIX B:

Beta Distribution

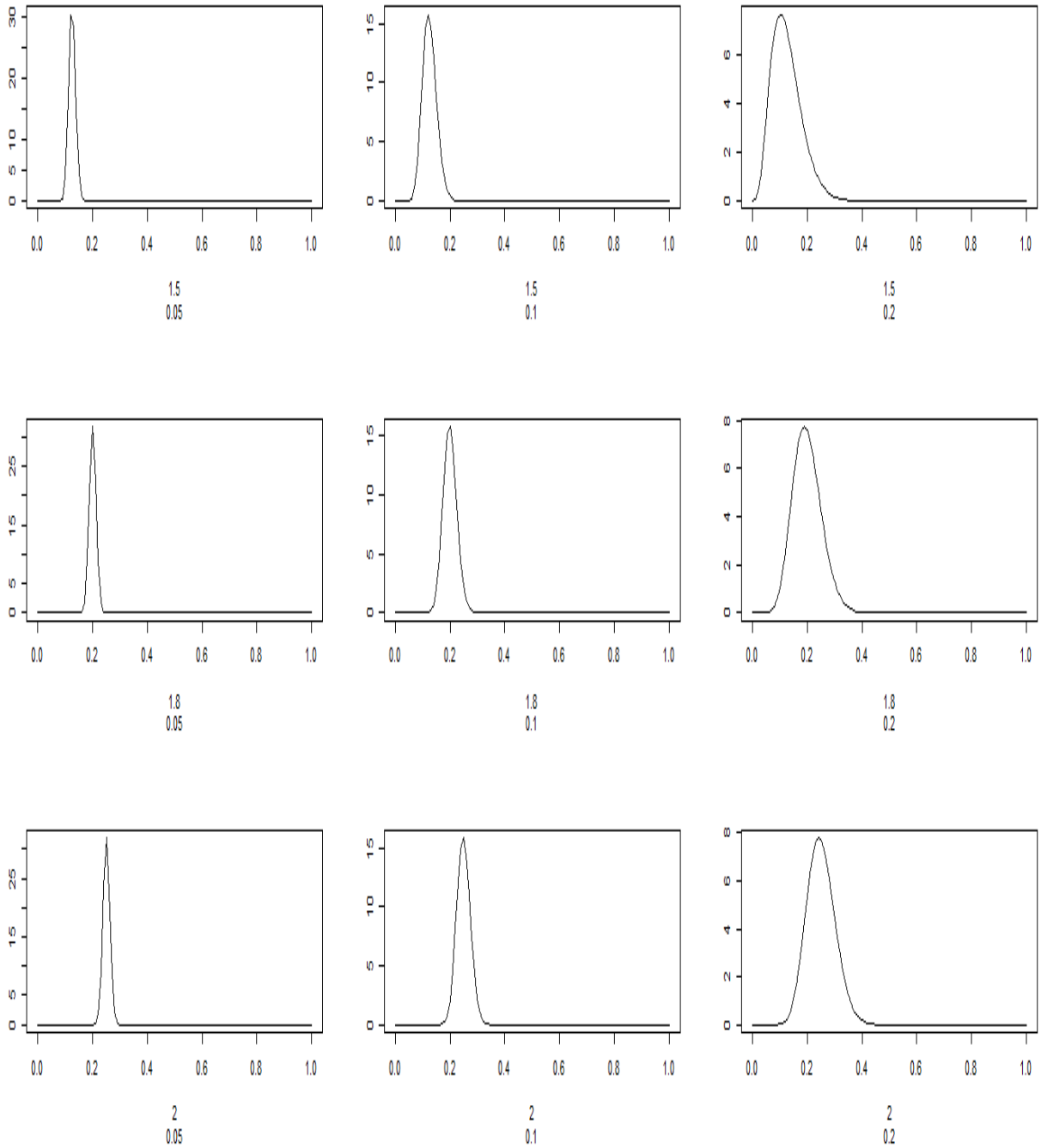

Figure B 1. Right-skewed beta distribution.

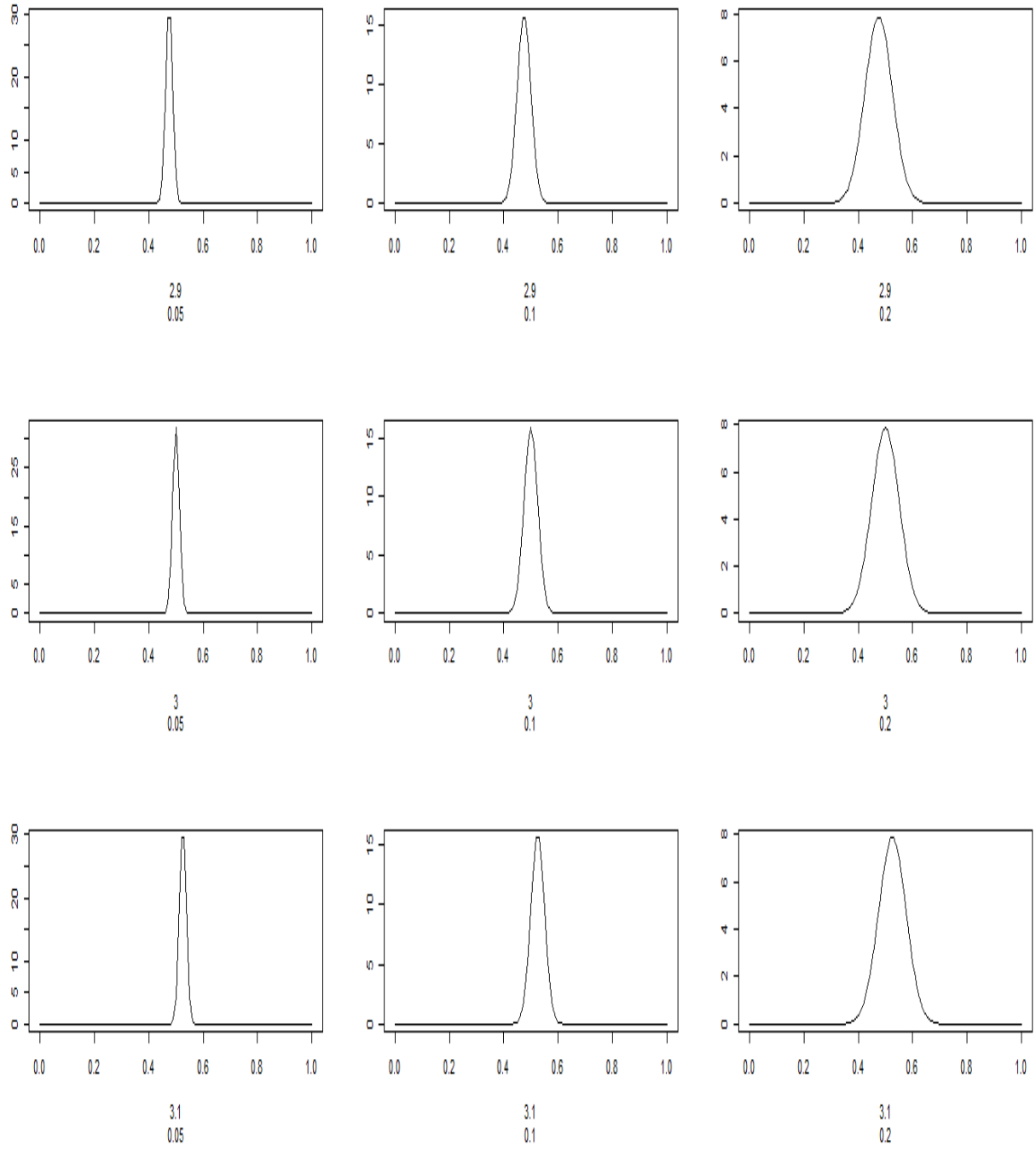

Figure B 2. Symmetrical beta distributions.

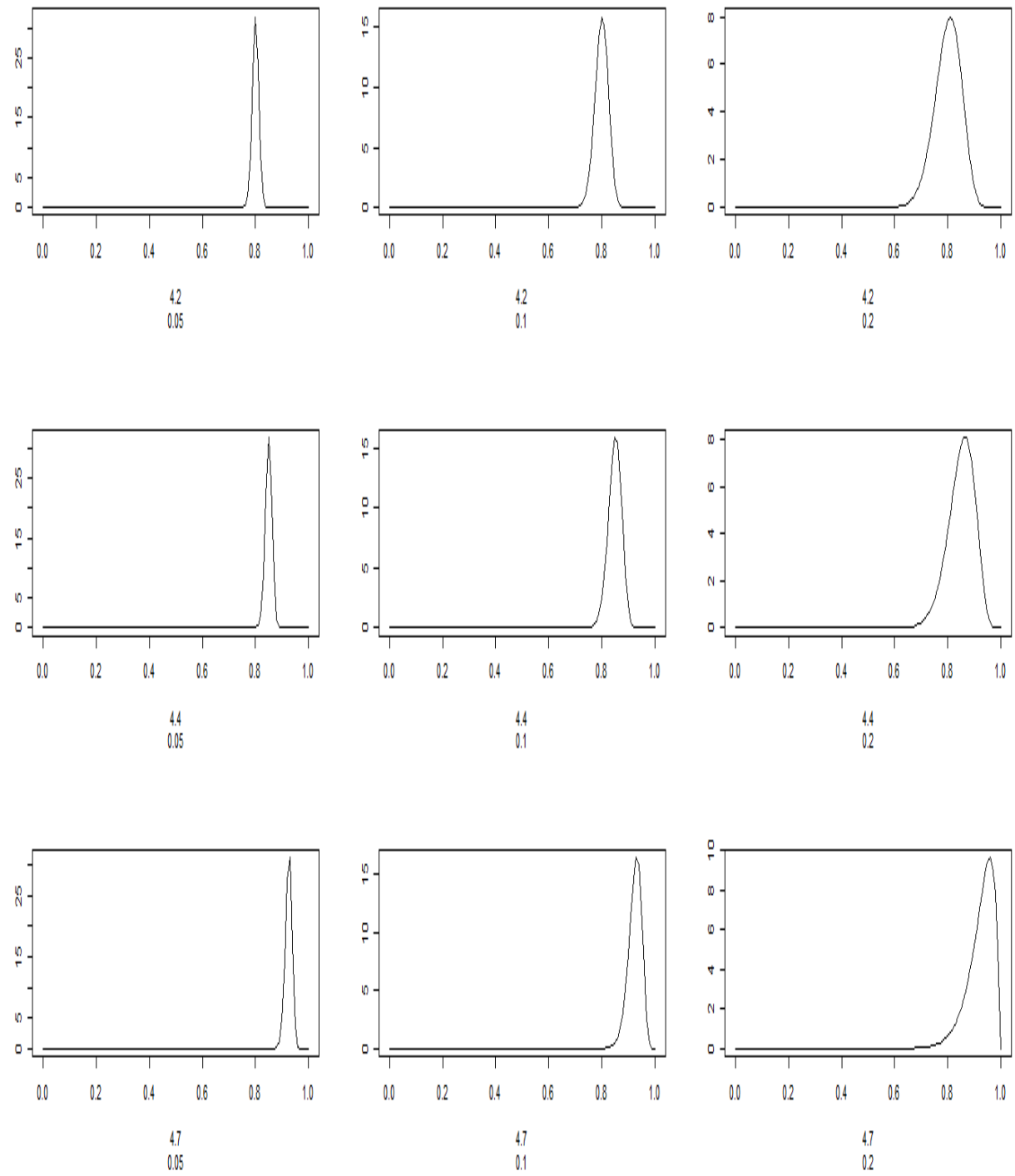

Figure B 3. Left-skewed beta distributions.

APPENDIX C:

Generalizability Results

Table C 1

Estimated Generalizability Coefficients for Varying Numbers of Items by Type of Design

Number of items	Items fixed design	Items random design
1	.867	.724
3	.890	.834
5	.895	.861
7	.898	.873
10	.899	.882

Table C 2

Estimated Generalizability Coefficients for Varying Numbers of Respondents by Type of Design

Number of Students	Items fixed design	Items random design
5	.613	.589
10	.760	.731
15	.826	.794
27	.895	.861
30	.905	.870
50	.941	.904
100	.969	.932

Table C 3

Estimated Generalizability Coefficients for Various Conditions by Type of Design

	Fixed item design							Random item design						
	Number of respondents							Number of respondents						
Number of items	5	10	15	27	30	50	100	5	10	15	27	30	50	100
1	.547	.707	.784	.867	.879	.924	.960	.456	.589	.653	.723	.732	.770	.800
3	.599	.749	.818	.890	.900	.937	.968	.562	.703	.767	.834	.843	.879	.907
5	.613	.760	.826	.895	.905	.941	.969	.589	.731	.794	.861	.870	.904	.932
7	.619	.764	.830	.898	.907	.942	.970	.602	.743	.807	.873	.882	.916	.943
10	.624	.768	.832	.899	.909	.943	.971	.611	.753	.816	.882	.891	.925	.952

APPENDIX D:

Simulation Tables

Table D 1

Coverage Properties of the 95% Confidence Interval of Different Sample Sizes (True Mean = 2.9)

		Logit		Resampling		BCA		Bayesboot		Z	
n	SD	CP	CE	CP	CE	CP	CE	CP	CE	CP	CE
5	0.050	0.945	0.005	0.828	0.122	0.836	0.114	0.801	0.149	0.896	0.054
5	0.100	0.948	0.002	0.824	0.126	0.836	0.114	0.805	0.145	0.890	0.060
5	0.200	0.955	0.005	0.848	0.102	0.855	0.095	0.828	0.122	0.890	0.060
10	0.050	0.948	0.002	0.888	0.062	0.884	0.066	0.870	0.080	0.931	0.019
10	0.100	0.951	0.001	0.890	0.060	0.890	0.060	0.874	0.076	0.929	0.021
10	0.200	0.955	0.005	0.914	0.036	0.908	0.042	0.901	0.049	0.937	0.013
20	0.050	0.942	0.008	0.909	0.041	0.910	0.040	0.900	0.050	0.955	0.005
20	0.100	0.947	0.003	0.922	0.028	0.920	0.030	0.914	0.036	0.943	0.007
20	0.200	0.945	0.005	0.910	0.040	0.917	0.033	0.915	0.035	0.939	0.011
50	0.050	0.945	0.005	0.933	0.017	0.924	0.026	0.921	0.029	0.956	0.006
50	0.100	0.947	0.003	0.936	0.014	0.925	0.025	0.935	0.015	0.946	0.004
50	0.200	0.945	0.005	0.934	0.016	0.935	0.015	0.931	0.019	0.957	0.007
100	0.050	0.945	0.005	0.936	0.014	0.919	0.031	0.925	0.025	0.958	0.008
100	0.100	0.948	0.002	0.944	0.006	0.937	0.013	0.937	0.013	0.944	0.006
100	0.200	0.950	0.000	0.946	0.004	0.938	0.012	0.943	0.007	0.943	0.007

Note. SD = standard deviation; CP = coverage probability; CE = coverage error.

Table D 2

Coverage Properties of the 95% Confidence Interval of Different Sample Sizes (True Mean = 3.0)

		Logit		Resampling		BCA		Bayesboot		Z	
n	SD	CP	CE	CP	CE	CP	CE	CP	CE	CP	CE
5	0.050	0.947	0.003	0.823	0.127	0.831	0.119	0.796	0.154	0.886	0.064
5	0.100	0.944	0.006	0.828	0.122	0.840	0.110	0.808	0.142	0.893	0.057
5	0.200	0.947	0.003	0.831	0.119	0.844	0.106	0.805	0.145	0.894	0.056
10	0.050	0.942	0.008	0.885	0.065	0.882	0.068	0.868	0.082	0.940	0.010
10	0.100	0.947	0.003	0.896	0.054	0.893	0.057	0.885	0.065	0.921	0.029
10	0.200	0.957	0.007	0.893	0.057	0.890	0.060	0.884	0.066	0.937	0.013
20	0.050	0.944	0.006	0.905	0.045	0.898	0.052	0.892	0.058	0.956	0.006
20	0.100	0.944	0.006	0.917	0.033	0.916	0.034	0.906	0.044	0.944	0.006
20	0.200	0.938	0.012	0.908	0.042	0.914	0.036	0.903	0.047	0.946	0.004
50	0.050	0.947	0.003	0.934	0.016	0.930	0.020	0.928	0.022	0.958	0.008
50	0.100	0.943	0.007	0.936	0.014	0.925	0.025	0.927	0.023	0.955	0.005
50	0.200	0.944	0.006	0.931	0.019	0.934	0.016	0.935	0.015	0.962	0.012
100	0.050	0.950	0.000	0.940	0.010	0.915	0.035	0.928	0.022	0.964	0.014
100	0.100	0.950	0.000	0.947	0.003	0.941	0.009	0.943	0.007	0.945	0.005
100	0.200	0.954	0.000	0.945	0.005	0.944	0.006	0.946	0.004	0.939	0.011

Note. SD = standard deviation; CP = coverage probability; CE = coverage error.

Table D 3

Coverage Properties of the 95% Confidence Interval of Different Sample Sizes (True Mean = 3.1)

		Logit		Resampling		BCA		Bayesboot		Z	
n	SD	CP	CE	CP	CE	CP	CE	CP	CE	CP	CE
5	0.050	0.945	0.005	0.826	0.124	0.833	0.117	0.801	0.149	0.896	0.054
5	0.100	0.948	0.002	0.823	0.127	0.842	0.108	0.805	0.145	0.890	0.060
5	0.200	0.955	0.005	0.848	0.102	0.858	0.092	0.828	0.122	0.890	0.060
10	0.050	0.948	0.002	0.889	0.061	0.885	0.065	0.870	0.080	0.931	0.019
10	0.100	0.951	0.001	0.891	0.059	0.891	0.059	0.874	0.076	0.929	0.021
10	0.200	0.955	0.005	0.916	0.034	0.907	0.043	0.901	0.049	0.937	0.013
20	0.050	0.942	0.008	0.910	0.040	0.910	0.040	0.900	0.050	0.955	0.005
20	0.100	0.947	0.003	0.923	0.027	0.920	0.030	0.914	0.036	0.943	0.007
20	0.200	0.945	0.005	0.913	0.037	0.916	0.034	0.914	0.036	0.939	0.011
50	0.050	0.945	0.005	0.935	0.015	0.924	0.026	0.921	0.029	0.956	0.006
50	0.100	0.947	0.003	0.938	0.012	0.925	0.025	0.935	0.015	0.946	0.004
50	0.200	0.945	0.005	0.935	0.015	0.935	0.015	0.931	0.019	0.957	0.007
100	0.050	0.945	0.005	0.935	0.015	0.919	0.031	0.925	0.025	0.958	0.008
100	0.100	0.948	0.002	0.945	0.005	0.937	0.013	0.937	0.013	0.944	0.006
100	0.200	0.950	0.000	0.946	0.004	0.938	0.012	0.943	0.007	0.943	0.007

Note. SD = standard deviation; CP = coverage probability; CE = coverage error.

Table D 4

Lower, Upper, and Width of the Confidence Interval of Different Sample Sizes (True Mean = 2.9)

		Logit			Resampling			BCA			Bayesboot			Z		
n	SD	L	U	W	L	U	W	L	U	W	L	U	W	L	U	W
5	0.05	2.841	2.957	0.117	2.864	2.934	0.070	2.860	2.935	0.076	2.866	2.932	0.066	2.858	2.941	0.083
5	0.10	2.781	3.014	0.234	2.827	2.967	0.140	2.819	2.971	0.152	2.831	2.964	0.132	2.816	2.981	0.165
5	0.20	2.659	3.133	0.474	2.753	3.037	0.284	2.736	3.044	0.308	2.762	3.030	0.268	2.730	3.062	0.332
10	0.05	2.865	2.934	0.069	2.872	2.928	0.056	2.871	2.929	0.058	2.873	2.927	0.054	2.870	2.930	0.061
10	0.10	2.830	2.969	0.139	2.843	2.956	0.113	2.842	2.958	0.116	2.845	2.954	0.109	2.839	2.960	0.121
10	0.20	2.758	3.040	0.282	2.785	3.012	0.227	2.784	3.016	0.232	2.790	3.009	0.219	2.777	3.021	0.244
20	0.05	2.877	2.923	0.046	2.879	2.921	0.042	2.879	2.921	0.042	2.879	2.920	0.041	2.878	2.921	0.043
20	0.10	2.853	2.946	0.093	2.858	2.941	0.084	2.858	2.942	0.084	2.858	2.941	0.082	2.856	2.943	0.087
20	0.20	2.806	2.993	0.187	2.815	2.983	0.168	2.815	2.984	0.169	2.817	2.982	0.166	2.811	2.985	0.174
50	0.05	2.886	2.914	0.028	2.886	2.913	0.027	2.886	2.914	0.027	2.886	2.914	0.027	2.886	2.914	0.028
50	0.10	2.871	2.928	0.057	2.872	2.927	0.054	2.872	2.927	0.055	2.872	2.927	0.055	2.872	2.927	0.055
50	0.20	2.842	2.957	0.115	2.845	2.954	0.109	2.844	2.955	0.110	2.845	2.955	0.110	2.844	2.955	0.111
100	0.05	2.890	2.910	0.020	2.890	2.910	0.019	2.890	2.910	0.019	2.890	2.910	0.020	2.890	2.910	0.020
100	0.10	2.880	2.920	0.040	2.881	2.920	0.039	2.881	2.920	0.039	2.881	2.920	0.039	2.880	2.919	0.039
100	0.20	2.860	2.941	0.080	2.862	2.939	0.078	2.862	2.940	0.078	2.861	2.940	0.079	2.860	2.939	0.079

Note. SD = standard deviation; L = lower limit; U = upper limit; W = average width.

Table D 5

Lower, Upper, and Width of the Confidence Interval of Different Sample Sizes (True Mean = 3.0)

		Logit			Resampling			BCA			Bayesboot			Z		
n	SD	L	U	W	L	U	W	L	U	W	L	U	W	L	U	W
5	0.05	2.940	3.057	0.116	2.964	3.034	0.070	2.960	3.035	0.076	2.966	3.032	0.066	2.958	3.040	0.083
5	0.10	2.881	3.114	0.233	2.928	3.067	0.140	2.919	3.071	0.152	2.932	3.064	0.132	2.915	3.081	0.165
5	0.20	2.759	3.230	0.471	2.853	3.136	0.282	2.836	3.142	0.306	2.862	3.128	0.266	2.831	3.162	0.332
10	0.05	2.965	3.034	0.069	2.972	3.028	0.056	2.971	3.029	0.058	2.973	3.027	0.054	2.970	3.030	0.061
10	0.10	2.930	3.069	0.139	2.943	3.056	0.113	2.942	3.058	0.115	2.945	3.054	0.109	2.939	3.060	0.121
10	0.20	2.858	3.140	0.281	2.885	3.112	0.227	2.883	3.116	0.232	2.890	3.109	0.219	2.878	3.120	0.243
20	0.05	2.977	3.023	0.046	2.979	3.021	0.042	2.979	3.021	0.042	2.979	3.020	0.041	2.978	3.021	0.043
20	0.10	2.953	3.046	0.093	2.958	3.041	0.083	2.958	3.042	0.084	2.959	3.041	0.083	2.956	3.043	0.087
20	0.20	2.906	3.093	0.187	2.915	3.083	0.168	2.915	3.084	0.169	2.917	3.082	0.166	2.911	3.085	0.174
50	0.05	2.986	3.014	0.028	2.986	3.013	0.027	2.986	3.014	0.027	2.986	3.014	0.027	2.986	3.014	0.028
50	0.10	2.971	3.028	0.057	2.973	3.027	0.054	2.972	3.027	0.055	2.973	3.027	0.055	2.972	3.027	0.055
50	0.20	2.942	3.057	0.115	2.945	3.054	0.109	2.945	3.055	0.110	2.945	3.055	0.110	2.944	3.054	0.111
100	0.05	2.990	3.010	0.020	2.990	3.010	0.019	2.990	3.010	0.019	2.990	3.010	0.020	2.990	3.010	0.020
100	0.10	2.980	3.020	0.040	2.981	3.020	0.039	2.981	3.020	0.039	2.981	3.020	0.039	2.980	3.019	0.039
100	0.20	2.961	3.041	0.080	2.962	3.040	0.078	2.962	3.040	0.078	2.961	3.040	0.079	2.960	3.039	0.079

Note. SD = standard deviation; L = lower limit; U = upper limit; W = average width.

Table D 6

Lower, Upper, and Width of the Confidence Interval of Different Sample Sizes (True Mean = 3.1)

		Logit			Resampling			BCA			Bayesboot			Z		
n	SD	L	U	W	L	U	W	L	U	W	L	U	W	L	U	W
5	0.05	3.043	3.159	0.117	3.066	3.136	0.070	3.061	3.137	0.076	3.068	3.134	0.066	3.059	3.142	0.083
5	0.10	2.986	3.219	0.234	3.032	3.172	0.140	3.023	3.175	0.152	3.036	3.169	0.132	3.019	3.184	0.165
5	0.20	2.868	3.341	0.473	2.962	3.246	0.284	2.942	3.250	0.308	2.970	3.238	0.268	2.938	3.270	0.332
10	0.05	3.066	3.135	0.069	3.072	3.128	0.056	3.071	3.129	0.058	3.073	3.127	0.054	3.070	3.130	0.061
10	0.10	3.031	3.170	0.139	3.044	3.157	0.113	3.042	3.158	0.116	3.046	3.155	0.109	3.040	3.161	0.121
10	0.20	2.960	3.242	0.282	2.987	3.214	0.227	2.984	3.216	0.232	2.991	3.210	0.219	2.979	3.223	0.243
20	0.05	3.077	3.123	0.046	3.079	3.121	0.042	3.079	3.121	0.042	3.080	3.121	0.041	3.079	3.122	0.043
20	0.10	3.054	3.147	0.093	3.059	3.142	0.083	3.058	3.142	0.084	3.059	3.142	0.082	3.057	3.144	0.087
20	0.20	3.007	3.194	0.187	3.017	3.184	0.168	3.016	3.185	0.169	3.018	3.183	0.165	3.015	3.189	0.174
50	0.05	3.086	3.114	0.028	3.086	3.114	0.027	3.086	3.114	0.027	3.086	3.114	0.027	3.086	3.114	0.028
50	0.10	3.072	3.129	0.057	3.073	3.127	0.054	3.073	3.128	0.055	3.073	3.127	0.055	3.073	3.128	0.055
50	0.20	3.043	3.158	0.115	3.045	3.155	0.109	3.045	3.156	0.110	3.046	3.155	0.110	3.045	3.156	0.111
100	0.05	3.090	3.110	0.020	3.090	3.109	0.019	3.090	3.110	0.019	3.090	3.110	0.020	3.090	3.110	0.020
100	0.10	3.080	3.120	0.040	3.080	3.119	0.039	3.080	3.119	0.039	3.080	3.119	0.039	3.081	3.120	0.039
100	0.20	3.059	3.139	0.080	3.060	3.138	0.078	3.060	3.138	0.078	3.060	3.139	0.079	3.061	3.140	0.079

Note. SD = standard deviation; L = lower limit; U = upper limit; W = average width.

Table D 7

Lower and Upper Error of the Confidence Interval of Different Sample Sizes (True Mean = 2.9)

		Logit		Resampling		BCA		Bayesboot		Z	
n	SD	LE	UE	LE	UE	LE	UE	LE	UE	LE	UE
5	0.050	0.021	0.034	0.076	0.096	0.056	0.094	0.084	0.106	0.051	0.053
5	0.100	0.024	0.028	0.077	0.099	0.069	0.090	0.081	0.107	0.055	0.055
5	0.200	0.017	0.028	0.065	0.087	0.058	0.085	0.078	0.089	0.053	0.057
10	0.050	0.025	0.027	0.058	0.054	0.054	0.048	0.060	0.057	0.034	0.035
10	0.100	0.023	0.026	0.056	0.054	0.053	0.048	0.060	0.060	0.035	0.036
10	0.200	0.021	0.024	0.041	0.045	0.046	0.043	0.050	0.048	0.031	0.032
20	0.050	0.027	0.031	0.040	0.051	0.038	0.042	0.038	0.046	0.020	0.025
20	0.100	0.023	0.030	0.037	0.041	0.034	0.040	0.032	0.042	0.025	0.032
20	0.200	0.026	0.029	0.040	0.050	0.039	0.043	0.039	0.044	0.030	0.031
50	0.050	0.027	0.028	0.032	0.035	0.027	0.027	0.027	0.029	0.024	0.020
50	0.100	0.022	0.031	0.025	0.039	0.026	0.036	0.026	0.034	0.020	0.034
50	0.200	0.024	0.031	0.031	0.035	0.028	0.032	0.031	0.033	0.023	0.020
100	0.050	0.027	0.028	0.032	0.032	0.028	0.026	0.020	0.026	0.027	0.015
100	0.100	0.028	0.024	0.030	0.026	0.031	0.021	0.027	0.019	0.032	0.024
100	0.200	0.023	0.027	0.026	0.028	0.026	0.029	0.024	0.027	0.029	0.028

Note. SD = standard deviation; LE = lower coverage error; UE = upper coverage error.

Table D 8

Lower and Upper Error of the Confidence Interval of Different Sample Sizes (True Mean = 3.0)

		Logit		Resampling		BCA		Bayesboot		Z	
n	SD	LE	UE	LE	UE	LE	UE	LE	UE	LE	UE
5	0.050	0.021	0.032	0.077	0.100	0.059	0.093	0.088	0.102	0.058	0.056
5	0.100	0.025	0.031	0.075	0.097	0.062	0.090	0.087	0.104	0.056	0.051
5	0.200	0.020	0.033	0.073	0.096	0.061	0.092	0.082	0.106	0.059	0.047
10	0.050	0.024	0.034	0.054	0.061	0.053	0.053	0.059	0.060	0.032	0.028
10	0.100	0.029	0.024	0.056	0.048	0.058	0.040	0.062	0.047	0.044	0.035
10	0.200	0.022	0.021	0.055	0.052	0.058	0.052	0.058	0.055	0.034	0.029
20	0.050	0.022	0.034	0.047	0.048	0.037	0.045	0.042	0.050	0.022	0.022
20	0.100	0.024	0.032	0.035	0.048	0.033	0.045	0.035	0.048	0.027	0.029
20	0.200	0.027	0.035	0.043	0.049	0.038	0.046	0.046	0.050	0.030	0.024
50	0.050	0.025	0.028	0.032	0.034	0.024	0.024	0.028	0.031	0.021	0.021
50	0.100	0.027	0.030	0.028	0.036	0.030	0.032	0.029	0.035	0.019	0.026
50	0.200	0.025	0.031	0.031	0.038	0.029	0.034	0.027	0.032	0.020	0.018
100	0.050	0.027	0.023	0.031	0.029	0.025	0.021	0.024	0.020	0.024	0.012
100	0.100	0.029	0.021	0.029	0.024	0.030	0.019	0.026	0.019	0.031	0.024
100	0.200	0.025	0.021	0.029	0.026	0.027	0.024	0.026	0.023	0.031	0.030

Note. SD = standard deviation; LE = lower coverage error; UE = upper coverage error.

Table D 9

Lower and Upper Error of the Confidence Interval of Different Sample Sizes (True Mean = 3.1)

		Logit		Resampling		BCA		Bayesboot		Z	
n	SD	LE	UE	LE	UE	LE	UE	LE	UE	LE	UE
5	0.050	0.034	0.021	0.096	0.078	0.082	0.073	0.106	0.084	0.053	0.051
5	0.100	0.028	0.024	0.098	0.079	0.078	0.076	0.107	0.081	0.055	0.055
5	0.200	0.028	0.017	0.087	0.065	0.072	0.067	0.089	0.078	0.057	0.053
10	0.050	0.027	0.025	0.053	0.058	0.048	0.054	0.057	0.060	0.035	0.034
10	0.100	0.026	0.023	0.053	0.056	0.048	0.053	0.060	0.060	0.036	0.035
10	0.200	0.024	0.021	0.042	0.042	0.043	0.046	0.048	0.050	0.032	0.031
20	0.050	0.031	0.027	0.049	0.041	0.042	0.038	0.046	0.038	0.025	0.020
20	0.100	0.030	0.023	0.040	0.037	0.040	0.034	0.042	0.032	0.032	0.025
20	0.200	0.029	0.026	0.047	0.040	0.043	0.039	0.044	0.039	0.031	0.030
50	0.050	0.028	0.027	0.032	0.033	0.027	0.027	0.029	0.027	0.020	0.024
50	0.100	0.031	0.022	0.037	0.025	0.036	0.026	0.034	0.026	0.034	0.020
50	0.200	0.031	0.024	0.034	0.031	0.032	0.028	0.033	0.031	0.020	0.023
100	0.050	0.028	0.027	0.032	0.033	0.026	0.028	0.026	0.020	0.015	0.027
100	0.100	0.024	0.028	0.025	0.030	0.021	0.031	0.019	0.027	0.024	0.032
100	0.200	0.027	0.023	0.026	0.028	0.029	0.026	0.027	0.024	0.028	0.029

Note. SD = standard deviation; LE = lower coverage error; UE = upper coverage error.

Table D 10

Coverage Properties of the 95% Confidence Interval of Different Sample Sizes (True Mean = 1.5)

		Logit		Resampling		BCA		Bayesboot		Z	
n	SD	CP	CE	CP	CE	CP	CE	CP	CE	CP	CE
5	0.050	0.946	0.004	0.818	0.132	0.828	0.122	0.798	0.152	0.881	0.069
5	0.100	0.944	0.006	0.823	0.127	0.839	0.111	0.800	0.150	0.897	0.053
5	0.200	0.939	0.011	0.808	0.142	0.822	0.128	0.791	0.159	0.873	0.077
10	0.050	0.947	0.003	0.891	0.059	0.887	0.063	0.874	0.076	0.910	0.040
10	0.100	0.948	0.002	0.894	0.056	0.894	0.056	0.882	0.068	0.920	0.030
10	0.200	0.951	0.001	0.898	0.052	0.897	0.053	0.885	0.065	0.904	0.046
20	0.050	0.939	0.011	0.906	0.044	0.905	0.045	0.891	0.059	0.938	0.012
20	0.100	0.946	0.004	0.911	0.039	0.912	0.038	0.903	0.047	0.938	0.012
20	0.200	0.938	0.012	0.903	0.047	0.912	0.038	0.904	0.046	0.936	0.014
50	0.050	0.939	0.011	0.922	0.028	0.907	0.043	0.911	0.039	0.945	0.005
50	0.100	0.960	0.010	0.949	0.001	0.947	0.003	0.944	0.006	0.943	0.007
50	0.200	0.948	0.002	0.928	0.022	0.930	0.020	0.925	0.025	0.944	0.006
100	0.050	0.954	0.004	0.950	0.000	0.937	0.013	0.940	0.010	0.954	0.004
100	0.100	0.946	0.004	0.942	0.008	0.931	0.019	0.939	0.011	0.948	0.002
100	0.200	0.954	0.004	0.940	0.010	0.944	0.006	0.943	0.007	0.954	0.004

Note. SD = standard deviation; CP = coverage probability; CE = coverage error.

Table D 11

Coverage Properties of the 95% Confidence Interval of Different Sample Sizes (True Mean = 1.8)

		Logit		Resampling		BCA		Bayesboot		Z	
n	SD	CP	CE	CP	CE	CP	CE	CP	CE	CP	CE
5	0.050	0.943	0.007	0.819	0.131	0.824	0.126	0.796	0.154	0.885	0.065
5	0.100	0.950	0.000	0.826	0.124	0.844	0.106	0.815	0.135	0.887	0.063
5	0.200	0.931	0.019	0.814	0.136	0.826	0.124	0.784	0.166	0.897	0.053
10	0.050	0.946	0.004	0.883	0.067	0.883	0.067	0.870	0.080	0.929	0.021
10	0.100	0.954	0.004	0.902	0.048	0.898	0.052	0.892	0.058	0.924	0.026
10	0.200	0.950	0.000	0.889	0.061	0.886	0.064	0.878	0.072	0.916	0.034
20	0.050	0.940	0.010	0.907	0.043	0.894	0.056	0.887	0.063	0.927	0.023
20	0.100	0.937	0.013	0.900	0.050	0.905	0.045	0.897	0.053	0.932	0.018
20	0.200	0.942	0.008	0.909	0.041	0.911	0.039	0.902	0.048	0.930	0.020
50	0.050	0.944	0.006	0.932	0.018	0.926	0.024	0.924	0.026	0.945	0.005
50	0.100	0.948	0.002	0.937	0.013	0.929	0.021	0.931	0.019	0.948	0.002
50	0.200	0.947	0.003	0.935	0.015	0.918	0.032	0.928	0.022	0.951	0.001
100	0.050	0.948	0.002	0.941	0.009	0.924	0.026	0.928	0.022	0.955	0.005
100	0.100	0.954	0.004	0.944	0.006	0.938	0.012	0.945	0.005	0.951	0.001
100	0.200	0.941	0.009	0.934	0.016	0.930	0.020	0.934	0.016	0.945	0.005

Note. SD = standard deviation; CP = coverage probability; CE = coverage error.

Table D 12

Coverage Properties of the 95% Confidence Interval of Different Sample Sizes (True Mean = 2.0)

		Logit		Resampling		BCA		Bayesboot		Z	
n	SD	CP	CE	CP	CE	CP	CE	CP	CE	CP	CE
5	0.050	0.956	0.006	0.829	0.121	0.840	0.110	0.802	0.148	0.900	0.050
5	0.100	0.945	0.005	0.823	0.127	0.836	0.114	0.795	0.155	0.884	0.066
5	0.200	0.949	0.001	0.830	0.120	0.844	0.106	0.807	0.143	0.891	0.059
10	0.050	0.950	0.000	0.898	0.052	0.896	0.054	0.881	0.069	0.928	0.022
10	0.100	0.948	0.002	0.911	0.039	0.908	0.042	0.893	0.057	0.916	0.034
10	0.200	0.943	0.007	0.895	0.055	0.903	0.047	0.887	0.063	0.921	0.029
20	0.050	0.943	0.007	0.913	0.037	0.906	0.044	0.896	0.054	0.932	0.018
20	0.100	0.939	0.011	0.906	0.044	0.907	0.043	0.903	0.047	0.936	0.014
20	0.200	0.940	0.010	0.919	0.031	0.924	0.026	0.917	0.033	0.931	0.019
50	0.050	0.939	0.011	0.928	0.022	0.919	0.031	0.925	0.025	0.956	0.006
50	0.100	0.947	0.003	0.936	0.014	0.941	0.009	0.931	0.019	0.945	0.005
50	0.200	0.943	0.007	0.922	0.028	0.933	0.017	0.925	0.025	0.946	0.004
100	0.050	0.943	0.007	0.935	0.015	0.920	0.030	0.926	0.024	0.951	0.001
100	0.100	0.950	0.000	0.943	0.007	0.936	0.014	0.943	0.007	0.933	0.017
100	0.200	0.947	0.003	0.932	0.018	0.931	0.019	0.934	0.016	0.937	0.013

Note. SD = standard deviation; CP = coverage probability; CE = coverage error.

Table D 13

Lower, Upper, and Width of the Confidence Interval of Different Sample Sizes (True Mean = 1.5)

		Logit			Resampling			BCA			Bayesboot			Z		
n	SD	L	U	W	L	U	W	L	U	W	L	U	W	L	U	W
5	0.05	1.443	1.560	0.117	1.464	1.534	0.070	1.460	1.536	0.076	1.466	1.532	0.066	1.458	1.541	0.082
5	0.10	1.389	1.626	0.238	1.427	1.568	0.141	1.420	1.573	0.152	1.432	1.565	0.133	1.415	1.583	0.168
5	0.20	1.283	1.802	0.520	1.348	1.646	0.298	1.338	1.662	0.323	1.359	1.640	0.281	1.320	1.676	0.356
10	0.05	1.466	1.535	0.070	1.472	1.528	0.056	1.472	1.529	0.058	1.473	1.527	0.054	1.469	1.530	0.061
10	0.10	1.432	1.573	0.142	1.443	1.557	0.114	1.444	1.561	0.117	1.446	1.556	0.110	1.438	1.561	0.123
10	0.20	1.360	1.667	0.307	1.380	1.622	0.242	1.385	1.636	0.251	1.387	1.621	0.234	1.369	1.631	0.261
20	0.05	1.477	1.523	0.046	1.479	1.521	0.042	1.479	1.521	0.042	1.479	1.521	0.041	1.478	1.521	0.044
20	0.10	1.454	1.548	0.094	1.457	1.542	0.084	1.458	1.544	0.086	1.459	1.542	0.083	1.455	1.543	0.088
20	0.20	1.403	1.607	0.204	1.410	1.590	0.180	1.415	1.599	0.184	1.414	1.592	0.178	1.406	1.594	0.188
50	0.05	1.486	1.514	0.029	1.486	1.514	0.027	1.486	1.514	0.028	1.486	1.514	0.027	1.486	1.514	0.028
50	0.10	1.471	1.530	0.058	1.472	1.528	0.055	1.473	1.529	0.056	1.473	1.528	0.056	1.472	1.528	0.056
50	0.20	1.440	1.566	0.126	1.442	1.560	0.118	1.444	1.564	0.120	1.443	1.562	0.118	1.440	1.560	0.120
100	0.05	1.490	1.510	0.020	1.491	1.510	0.019	1.491	1.510	0.020	1.490	1.510	0.020	1.490	1.510	0.020
100	0.10	1.480	1.521	0.041	1.481	1.520	0.039	1.481	1.521	0.040	1.481	1.521	0.040	1.480	1.520	0.040
100	0.20	1.458	1.546	0.088	1.459	1.543	0.084	1.460	1.546	0.085	1.459	1.545	0.086	1.458	1.543	0.085

Note. SD = standard deviation; L = lower limit; U = upper limit; W = average width.

Table D 14

Lower, Upper, and Width of the Confidence Interval of Different Sample Sizes (True Mean = 1.8)

		Logit			Resampling			BCA			Bayesboot			Z		
n	SD	L	U	W	L	U	W	L	U	W	L	U	W	L	U	W
5	0.05	1.742	1.858	0.117	1.764	1.834	0.070	1.760	1.835	0.076	1.766	1.832	0.066	1.758	1.841	0.083
5	0.10	1.686	1.920	0.234	1.728	1.868	0.140	1.720	1.872	0.152	1.732	1.864	0.132	1.716	1.881	0.165
5	0.20	1.578	2.059	0.481	1.653	1.937	0.284	1.640	1.946	0.306	1.662	1.930	0.268	1.628	1.968	0.340
10	0.05	1.765	1.835	0.070	1.772	1.828	0.056	1.771	1.829	0.058	1.773	1.827	0.054	1.769	1.830	0.061
10	0.10	1.731	1.872	0.140	1.743	1.857	0.113	1.743	1.859	0.116	1.746	1.855	0.109	1.739	1.861	0.122
10	0.20	1.662	1.951	0.289	1.684	1.915	0.231	1.686	1.923	0.238	1.690	1.913	0.223	1.675	1.924	0.248
20	0.05	1.777	1.823	0.046	1.779	1.821	0.042	1.779	1.821	0.042	1.779	1.820	0.041	1.778	1.821	0.043
20	0.10	1.754	1.847	0.093	1.758	1.842	0.084	1.758	1.843	0.085	1.759	1.841	0.083	1.756	1.843	0.087
20	0.20	1.705	1.898	0.193	1.713	1.885	0.171	1.715	1.889	0.174	1.716	1.885	0.169	1.710	1.888	0.178
50	0.05	1.786	1.814	0.029	1.786	1.814	0.027	1.786	1.814	0.027	1.786	1.814	0.027	1.786	1.814	0.028
50	0.10	1.772	1.829	0.058	1.773	1.828	0.055	1.773	1.828	0.055	1.773	1.828	0.055	1.772	1.828	0.055
50	0.20	1.742	1.861	0.119	1.744	1.856	0.112	1.745	1.858	0.113	1.745	1.857	0.113	1.743	1.856	0.113
100	0.05	1.790	1.810	0.020	1.791	1.810	0.019	1.790	1.810	0.020	1.790	1.810	0.020	1.790	1.810	0.020
100	0.10	1.781	1.821	0.040	1.781	1.820	0.039	1.781	1.820	0.039	1.781	1.820	0.040	1.780	1.820	0.039
100	0.20	1.760	1.843	0.083	1.761	1.841	0.080	1.762	1.842	0.080	1.761	1.842	0.081	1.760	1.840	0.081

Note. SD = standard deviation; L = lower limit; U = upper limit; W = average width.

Table D 15

Lower, Upper, and Width of the Confidence Interval of Different Sample Sizes (True Mean = 2.0)

		Logit			Resampling			BCA			Bayesboot			Z		
n	SD	L	U	W	L	U	W	L	U	W	L	U	W	L	U	W
5	0.05	1.941	2.058	0.117	1.964	2.034	0.070	1.960	2.035	0.076	1.966	2.032	0.066	1.958	2.041	0.083
5	0.10	1.885	2.118	0.234	1.928	2.067	0.140	1.920	2.071	0.152	1.932	2.064	0.132	1.916	2.081	0.165
5	0.20	1.773	2.251	0.477	1.853	2.136	0.284	1.838	2.145	0.307	1.862	2.130	0.268	1.829	2.168	0.339
10	0.05	1.965	2.035	0.070	1.971	2.028	0.056	1.971	2.029	0.058	1.973	2.027	0.054	1.970	2.030	0.061
10	0.10	1.931	2.071	0.140	1.943	2.056	0.113	1.943	2.059	0.116	1.945	2.055	0.110	1.939	2.060	0.121
10	0.20	1.863	2.147	0.284	1.886	2.114	0.228	1.887	2.121	0.234	1.891	2.112	0.220	1.876	2.123	0.247
20	0.05	1.977	2.023	0.046	1.979	2.021	0.042	1.979	2.021	0.042	1.979	2.020	0.041	1.978	2.021	0.043
20	0.10	1.954	2.047	0.093	1.958	2.041	0.083	1.958	2.043	0.085	1.959	2.041	0.083	1.956	2.043	0.087
20	0.20	1.907	2.097	0.190	1.915	2.085	0.170	1.916	2.088	0.172	1.917	2.085	0.168	1.911	2.088	0.177
50	0.05	1.986	2.014	0.029	1.986	2.014	0.027	1.986	2.014	0.028	1.986	2.014	0.027	1.986	2.014	0.028
50	0.10	1.971	2.029	0.057	1.973	2.027	0.055	1.973	2.028	0.055	1.973	2.028	0.055	1.972	2.027	0.055
50	0.20	1.942	2.060	0.118	1.945	2.056	0.111	1.946	2.058	0.112	1.945	2.057	0.111	1.944	2.056	0.112
100	0.05	1.990	2.010	0.020	1.990	2.010	0.019	1.991	2.010	0.019	1.990	2.010	0.020	1.990	2.010	0.020
100	0.10	1.981	2.021	0.040	1.981	2.020	0.039	1.981	2.020	0.039	1.981	2.020	0.039	1.980	2.020	0.039
100	0.20	1.961	2.043	0.082	1.962	2.041	0.079	1.962	2.042	0.079	1.962	2.042	0.080	1.960	2.040	0.080

Note. SD = standard deviation; L = lower limit; U = upper limit; W = average width.

Table D 16

Lower and Upper Error of the Confidence Interval of Different Sample Sizes (True Mean = 1.5)

		Logit		Resampling		BCA		Bayesboot		Z	
n	SD	LE	UE	LE	UE	LE	UE	LE	UE	LE	UE
5	0.050	0.020	0.034	0.075	0.107	0.061	0.096	0.081	0.106	0.056	0.063
5	0.100	0.016	0.040	0.065	0.112	0.051	0.108	0.077	0.118	0.046	0.057
5	0.200	0.013	0.048	0.048	0.144	0.041	0.133	0.054	0.151	0.030	0.097
10	0.050	0.027	0.026	0.049	0.060	0.049	0.052	0.053	0.061	0.044	0.046
10	0.100	0.022	0.030	0.038	0.068	0.043	0.061	0.045	0.068	0.041	0.039
10	0.200	0.013	0.036	0.029	0.073	0.035	0.066	0.034	0.078	0.026	0.070
20	0.050	0.027	0.034	0.039	0.055	0.034	0.050	0.039	0.051	0.026	0.036
20	0.100	0.023	0.031	0.037	0.052	0.041	0.040	0.041	0.049	0.023	0.039
20	0.200	0.019	0.043	0.027	0.070	0.033	0.053	0.029	0.064	0.023	0.041
50	0.050	0.032	0.029	0.036	0.042	0.035	0.028	0.033	0.031	0.030	0.025
50	0.100	0.014	0.026	0.019	0.032	0.019	0.027	0.018	0.027	0.024	0.033
50	0.200	0.021	0.031	0.027	0.045	0.030	0.036	0.027	0.041	0.021	0.035
100	0.050	0.025	0.021	0.026	0.024	0.022	0.018	0.023	0.017	0.027	0.019
100	0.100	0.027	0.027	0.030	0.028	0.031	0.025	0.027	0.024	0.028	0.024
100	0.200	0.019	0.027	0.023	0.037	0.025	0.024	0.025	0.027	0.023	0.023

Note. SD = standard deviation; LE = lower coverage error; UE = upper coverage error.

Table D 17

Lower and Upper Error of the Confidence Interval of Different Sample Sizes (True Mean = 1.8)

		Logit		Resampling		BCA		Bayesboot		Z	
n	SD	LE	UE	LE	UE	LE	UE	LE	UE	LE	UE
5	0.050	0.021	0.036	0.073	0.108	0.062	0.101	0.081	0.110	0.057	0.058
5	0.100	0.014	0.036	0.070	0.104	0.057	0.093	0.077	0.106	0.048	0.065
5	0.200	0.020	0.049	0.067	0.119	0.058	0.113	0.085	0.129	0.037	0.066
10	0.050	0.025	0.029	0.057	0.060	0.054	0.055	0.053	0.063	0.037	0.034
10	0.100	0.020	0.026	0.041	0.057	0.041	0.052	0.044	0.059	0.040	0.036
10	0.200	0.022	0.028	0.044	0.067	0.053	0.058	0.049	0.072	0.035	0.049
20	0.050	0.029	0.031	0.044	0.049	0.043	0.045	0.043	0.051	0.028	0.045
20	0.100	0.028	0.035	0.043	0.057	0.044	0.046	0.042	0.047	0.029	0.039
20	0.200	0.031	0.027	0.044	0.047	0.046	0.038	0.047	0.045	0.031	0.039
50	0.050	0.030	0.026	0.034	0.034	0.030	0.026	0.032	0.024	0.029	0.026
50	0.100	0.026	0.026	0.028	0.035	0.030	0.029	0.028	0.029	0.023	0.029
50	0.200	0.020	0.033	0.023	0.042	0.030	0.043	0.025	0.043	0.024	0.025
100	0.050	0.030	0.022	0.032	0.027	0.027	0.025	0.026	0.020	0.027	0.018
100	0.100	0.028	0.018	0.033	0.023	0.031	0.016	0.030	0.016	0.029	0.020
100	0.200	0.023	0.036	0.025	0.041	0.029	0.039	0.025	0.040	0.032	0.023

Note. SD = standard deviation; LE = lower coverage error; UE = upper coverage error.

Table D 18

Lower and Upper Error of the Confidence Interval of Different Sample Sizes (True Mean = 2.0)

		Logit		Resampling		BCA		Bayesboot		Z	
n	SD	LE	UE	LE	UE	LE	UE	LE	UE	LE	UE
5	0.050	0.014	0.030	0.071	0.100	0.054	0.092	0.079	0.106	0.047	0.053
5	0.100	0.023	0.032	0.069	0.108	0.058	0.102	0.076	0.120	0.050	0.066
5	0.200	0.014	0.037	0.060	0.110	0.051	0.103	0.072	0.119	0.039	0.070
10	0.050	0.026	0.024	0.051	0.051	0.051	0.041	0.057	0.048	0.038	0.034
10	0.100	0.022	0.030	0.041	0.048	0.042	0.044	0.043	0.055	0.042	0.042
10	0.200	0.024	0.033	0.047	0.058	0.048	0.049	0.050	0.060	0.039	0.040
20	0.050	0.029	0.028	0.043	0.044	0.036	0.037	0.042	0.042	0.028	0.040
20	0.100	0.028	0.033	0.041	0.053	0.041	0.048	0.036	0.051	0.030	0.034
20	0.200	0.030	0.030	0.037	0.044	0.038	0.036	0.036	0.042	0.030	0.039
50	0.050	0.027	0.034	0.035	0.037	0.029	0.030	0.028	0.033	0.022	0.022
50	0.100	0.030	0.023	0.038	0.026	0.035	0.022	0.034	0.023	0.025	0.030
50	0.200	0.029	0.028	0.038	0.040	0.032	0.032	0.035	0.038	0.018	0.036
100	0.050	0.028	0.029	0.030	0.035	0.022	0.026	0.026	0.020	0.028	0.021
100	0.100	0.028	0.022	0.031	0.026	0.030	0.021	0.028	0.018	0.039	0.028
100	0.200	0.029	0.024	0.035	0.033	0.034	0.027	0.032	0.029	0.032	0.031

Note. SD = standard deviation; LE = lower coverage error; UE = upper coverage error.

Table D 19

Coverage Properties of the 95% Confidence Interval of Different Sample Sizes (True Mean = 4.2)

		Logit		Resampling		Bca		Bayesboot		Z	
n	SD	CP	CE	CP	CE	CP	CE	CP	CE	CP	CE
5	0.050	0.943	0.007	0.818	0.132	0.826	0.124	0.796	0.154	0.885	0.065
5	0.100	0.948	0.002	0.822	0.128	0.839	0.111	0.803	0.147	0.882	0.068
5	0.200	0.953	0.003	0.817	0.133	0.836	0.114	0.802	0.148	0.896	0.054
10	0.050	0.946	0.004	0.883	0.067	0.884	0.066	0.870	0.080	0.929	0.021
10	0.100	0.952	0.002	0.898	0.052	0.900	0.050	0.890	0.060	0.920	0.030
10	0.200	0.945	0.005	0.894	0.056	0.899	0.051	0.885	0.065	0.920	0.030
20	0.050	0.938	0.012	0.907	0.043	0.898	0.052	0.890	0.060	0.930	0.020
20	0.100	0.934	0.016	0.903	0.047	0.905	0.045	0.893	0.057	0.934	0.016
20	0.200	0.948	0.002	0.919	0.031	0.918	0.032	0.917	0.033	0.935	0.015
50	0.050	0.949	0.001	0.940	0.010	0.927	0.023	0.930	0.020	0.949	0.001
50	0.100	0.943	0.007	0.933	0.017	0.928	0.022	0.928	0.022	0.949	0.001
50	0.200	0.956	0.006	0.936	0.014	0.934	0.016	0.936	0.014	0.950	0.000
100	0.050	0.949	0.001	0.946	0.004	0.931	0.019	0.931	0.019	0.955	0.005
100	0.100	0.956	0.006	0.950	0.000	0.948	0.002	0.942	0.008	0.954	0.004
100	0.200	0.943	0.007	0.935	0.015	0.928	0.022	0.937	0.013	0.944	0.006

Note. SD = standard deviation; CP = coverage probability; CE = coverage error.

Table D 20

Coverage Properties of the 95% Confidence Interval of Different Sample Sizes (True Mean = 4.4)

		Logit		Resampling		Bca		Bayesboot		Z	
n	SD	CP	CE	CP	CE	CP	CE	CP	CE	CP	CE
5	0.050	0.942	0.008	0.823	0.127	0.838	0.112	0.796	0.154	0.889	0.061
5	0.100	0.949	0.001	0.818	0.132	0.829	0.121	0.795	0.155	0.881	0.069
5	0.200	0.941	0.009	0.813	0.137	0.818	0.132	0.788	0.162	0.886	0.064
10	0.050	0.955	0.005	0.900	0.050	0.899	0.051	0.885	0.065	0.922	0.028
10	0.100	0.951	0.001	0.902	0.048	0.899	0.051	0.892	0.058	0.917	0.033
10	0.200	0.944	0.006	0.896	0.054	0.889	0.061	0.870	0.080	0.913	0.037
20	0.050	0.956	0.006	0.929	0.021	0.927	0.023	0.921	0.029	0.937	0.013
20	0.100	0.939	0.011	0.915	0.035	0.909	0.041	0.905	0.045	0.941	0.009
20	0.200	0.939	0.011	0.894	0.056	0.892	0.058	0.886	0.064	0.930	0.020
50	0.050	0.936	0.014	0.930	0.020	0.925	0.025	0.920	0.030	0.952	0.002
50	0.100	0.934	0.016	0.923	0.027	0.917	0.033	0.918	0.032	0.947	0.003
50	0.200	0.962	0.012	0.945	0.005	0.937	0.013	0.940	0.010	0.938	0.012
100	0.050	0.941	0.009	0.935	0.015	0.921	0.029	0.924	0.026	0.955	0.005
100	0.100	0.954	0.004	0.946	0.004	0.940	0.010	0.939	0.011	0.945	0.005
100	0.200	0.943	0.007	0.934	0.016	0.932	0.018	0.935	0.015	0.952	0.002

Note. SD = standard deviation; CP = coverage probability; CE = coverage error.

Table D 21

Coverage Properties of the 95% Confidence Interval of Different Sample Sizes (True Mean = 4.7)

		Logit		Resampling		Bca		Bayesboot		Z	
n	SD	CP	CE	CP	CE	CP	CE	CP	CE	CP	CE
5	0.050	0.947	0.003	0.824	0.126	0.835	0.115	0.793	0.157	0.893	0.057
5	0.100	0.936	0.014	0.825	0.125	0.836	0.114	0.798	0.152	0.867	0.083
5	0.200	0.926	0.024	0.806	0.144	0.833	0.117	0.788	0.162	0.869	0.081
10	0.050	0.942	0.008	0.905	0.045	0.891	0.059	0.882	0.068	0.914	0.036
10	0.100	0.943	0.007	0.878	0.072	0.881	0.069	0.859	0.091	0.906	0.044
10	0.200	0.942	0.008	0.887	0.063	0.893	0.057	0.882	0.068	0.887	0.063
20	0.050	0.939	0.011	0.913	0.037	0.897	0.053	0.893	0.057	0.931	0.019
20	0.100	0.941	0.009	0.907	0.043	0.904	0.046	0.899	0.051	0.922	0.028
20	0.200	0.938	0.012	0.915	0.035	0.921	0.029	0.912	0.038	0.921	0.029
50	0.050	0.941	0.009	0.927	0.023	0.915	0.035	0.916	0.034	0.954	0.004
50	0.100	0.949	0.001	0.940	0.010	0.931	0.019	0.933	0.017	0.934	0.016
50	0.200	0.930	0.020	0.921	0.029	0.921	0.029	0.920	0.030	0.946	0.004
100	0.050	0.944	0.006	0.939	0.011	0.930	0.020	0.932	0.018	0.939	0.011
100	0.100	0.938	0.012	0.934	0.016	0.927	0.023	0.930	0.020	0.954	0.004
100	0.200	0.951	0.001	0.951	0.001	0.948	0.002	0.954	0.004	0.956	0.006

Note. SD = standard deviation; CP = coverage probability; CE = coverage error.

Table D 22

Lower, Upper, and Width of the Confidence Interval of Different Sample Sizes (True Mean = 4.2)

		Logit			Resampling			BCA			Bayesboot			Z		
n	SD	L	U	W	L	U	W	L	U	W	L	U	W	L	U	W
5	0.05	4.142	4.258	0.116	4.166	4.236	0.070	4.161	4.237	0.076	4.168	4.234	0.066	4.159	4.242	0.083
5	0.10	4.081	4.313	0.232	4.132	4.272	0.139	4.122	4.273	0.151	4.136	4.267	0.131	4.119	4.283	0.164
5	0.20	3.948	4.418	0.470	4.066	4.343	0.277	4.043	4.346	0.302	4.073	4.335	0.262	4.037	4.367	0.330
10	0.05	4.165	4.235	0.069	4.172	4.228	0.056	4.171	4.229	0.058	4.173	4.227	0.054	4.170	4.230	0.060
10	0.10	4.129	4.268	0.139	4.144	4.256	0.112	4.141	4.256	0.115	4.145	4.254	0.108	4.140	4.261	0.121
10	0.20	4.054	4.335	0.281	4.088	4.312	0.224	4.081	4.311	0.230	4.091	4.307	0.217	4.080	4.321	0.241
20	0.05	4.177	4.223	0.046	4.179	4.221	0.042	4.179	4.221	0.042	4.180	4.221	0.041	4.179	4.222	0.043
20	0.10	4.154	4.246	0.092	4.158	4.241	0.083	4.157	4.241	0.084	4.159	4.241	0.082	4.157	4.244	0.087
20	0.20	4.105	4.292	0.187	4.117	4.284	0.167	4.114	4.282	0.168	4.118	4.282	0.164	4.115	4.287	0.172
50	0.05	4.186	4.214	0.028	4.186	4.214	0.027	4.186	4.214	0.027	4.186	4.214	0.027	4.186	4.214	0.028
50	0.10	4.171	4.228	0.057	4.172	4.227	0.054	4.172	4.227	0.055	4.172	4.227	0.055	4.173	4.228	0.055
50	0.20	4.141	4.257	0.116	4.145	4.254	0.109	4.143	4.253	0.110	4.145	4.254	0.109	4.146	4.255	0.110
100	0.05	4.190	4.210	0.020	4.190	4.209	0.019	4.190	4.209	0.019	4.190	4.210	0.020	4.190	4.210	0.020
100	0.10	4.179	4.219	0.040	4.180	4.219	0.039	4.180	4.219	0.039	4.180	4.219	0.039	4.180	4.219	0.039
100	0.20	4.158	4.239	0.081	4.160	4.237	0.077	4.159	4.237	0.078	4.159	4.238	0.078	4.161	4.239	0.078

Note. SD = standard deviation; L = lower limit; U = upper limit; W = average width.

Table D 23

Lower, Upper, and Width of the Confidence Interval of Different Sample Sizes (True Mean = 4.4)

		Logit			Resampling			BCA			Bayesboot			Z		
n	SD	L	U	W	L	U	W	L	U	W	L	U	W	L	U	W
5	0.05	4.341	4.458	0.116	4.366	4.436	0.070	4.361	4.437	0.076	4.368	4.434	0.066	4.359	4.442	0.083
5	0.10	4.278	4.511	0.233	4.333	4.471	0.139	4.322	4.472	0.151	4.336	4.467	0.131	4.319	4.483	0.164
5	0.20	4.134	4.608	0.474	4.264	4.541	0.277	4.239	4.541	0.302	4.270	4.531	0.261	4.238	4.567	0.329
10	0.05	4.365	4.434	0.070	4.372	4.428	0.056	4.371	4.428	0.058	4.373	4.427	0.054	4.370	4.430	0.060
10	0.10	4.328	4.468	0.139	4.344	4.456	0.112	4.340	4.456	0.115	4.345	4.454	0.108	4.340	4.461	0.121
10	0.20	4.250	4.531	0.280	4.287	4.510	0.223	4.278	4.507	0.230	4.289	4.505	0.215	4.279	4.520	0.240
20	0.05	4.377	4.423	0.046	4.379	4.421	0.042	4.379	4.421	0.042	4.379	4.421	0.041	4.379	4.422	0.043
20	0.10	4.353	4.446	0.092	4.358	4.441	0.083	4.357	4.441	0.084	4.359	4.441	0.082	4.357	4.444	0.086
20	0.20	4.304	4.490	0.186	4.317	4.482	0.165	4.312	4.480	0.168	4.317	4.480	0.163	4.314	4.487	0.173
50	0.05	4.386	4.414	0.028	4.386	4.414	0.027	4.386	4.414	0.027	4.386	4.414	0.027	4.386	4.414	0.028
50	0.10	4.371	4.428	0.057	4.372	4.427	0.054	4.372	4.427	0.055	4.372	4.427	0.054	4.373	4.428	0.055
50	0.20	4.340	4.456	0.116	4.345	4.454	0.109	4.343	4.452	0.110	4.344	4.453	0.109	4.345	4.455	0.110
100	0.05	4.390	4.410	0.020	4.390	4.409	0.019	4.390	4.409	0.019	4.390	4.410	0.020	4.390	4.410	0.020
100	0.10	4.379	4.419	0.040	4.380	4.419	0.039	4.380	4.419	0.039	4.380	4.419	0.039	4.380	4.420	0.039
100	0.20	4.358	4.439	0.081	4.360	4.437	0.077	4.359	4.437	0.078	4.359	4.438	0.078	4.361	4.439	0.078

Note. SD = standard deviation; L = lower limit; U = upper limit; W = average width.

Table D 24

Lower, Upper, and Width of the Confidence Interval of Different Sample Sizes (True Mean = 4.7)

		Logit			Resampling			BCA			Bayesboot			Z		
n	SD	L	U	W	L	U	W	L	U	W	L	U	W	L	U	W
5	0.05	4.640	4.755	0.116	4.666	4.735	0.069	4.661	4.736	0.075	4.668	4.733	0.065	4.660	4.742	0.082
5	0.10	4.570	4.804	0.234	4.633	4.770	0.138	4.620	4.770	0.150	4.636	4.765	0.130	4.620	4.782	0.162
5	0.20	4.399	4.880	0.481	4.561	4.831	0.270	4.527	4.827	0.300	4.566	4.820	0.254	4.538	4.860	0.322
10	0.05	4.664	4.734	0.069	4.672	4.728	0.056	4.670	4.728	0.057	4.673	4.727	0.054	4.670	4.730	0.060
10	0.10	4.627	4.765	0.138	4.644	4.755	0.111	4.639	4.753	0.114	4.645	4.752	0.107	4.640	4.760	0.119
10	0.20	4.540	4.818	0.278	4.583	4.804	0.221	4.565	4.797	0.232	4.584	4.796	0.213	4.582	4.817	0.235
20	0.05	4.677	4.723	0.046	4.679	4.721	0.041	4.678	4.720	0.042	4.679	4.720	0.041	4.679	4.722	0.043
20	0.10	4.653	4.745	0.092	4.659	4.741	0.082	4.656	4.740	0.084	4.658	4.740	0.082	4.657	4.743	0.086
20	0.20	4.601	4.783	0.182	4.615	4.779	0.165	4.604	4.774	0.170	4.613	4.775	0.162	4.615	4.786	0.171
50	0.05	4.685	4.714	0.028	4.686	4.713	0.027	4.686	4.713	0.027	4.686	4.713	0.027	4.686	4.714	0.027
50	0.10	4.671	4.727	0.057	4.672	4.726	0.054	4.671	4.726	0.055	4.672	4.726	0.054	4.673	4.728	0.055
50	0.20	4.640	4.752	0.112	4.643	4.752	0.109	4.638	4.749	0.111	4.642	4.750	0.109	4.645	4.754	0.110
100	0.05	4.690	4.710	0.020	4.690	4.709	0.019	4.690	4.709	0.019	4.690	4.709	0.020	4.690	4.710	0.020
100	0.10	4.679	4.719	0.040	4.680	4.719	0.039	4.679	4.718	0.039	4.680	4.719	0.039	4.681	4.719	0.039
100	0.20	4.658	4.736	0.078	4.659	4.736	0.078	4.656	4.735	0.079	4.657	4.736	0.079	4.660	4.738	0.078

Note. SD = standard deviation; L = lower limit; U = upper limit; W = average width.

Table D 25

Lower and Upper Error of the Confidence Interval of Different Sample Sizes (True Mean = 4.2)

		Logit		Resampling		BCA		Bayesboot		Z	
n	SD	LE	UE	LE	UE	LE	UE	LE	UE	LE	UE
5	0.050	0.036	0.021	0.106	0.076	0.084	0.076	0.110	0.081	0.058	0.057
5	0.100	0.038	0.014	0.106	0.072	0.087	0.066	0.117	0.077	0.065	0.053
5	0.200	0.037	0.010	0.119	0.064	0.101	0.058	0.126	0.068	0.063	0.041
10	0.050	0.029	0.025	0.060	0.057	0.055	0.055	0.063	0.053	0.034	0.037
10	0.100	0.029	0.019	0.057	0.045	0.053	0.037	0.061	0.047	0.043	0.037
10	0.200	0.031	0.024	0.066	0.040	0.058	0.041	0.070	0.038	0.051	0.029
20	0.050	0.032	0.030	0.050	0.043	0.041	0.043	0.047	0.041	0.043	0.027
20	0.100	0.034	0.032	0.053	0.044	0.044	0.042	0.052	0.045	0.033	0.033
20	0.200	0.030	0.022	0.043	0.038	0.040	0.036	0.043	0.038	0.036	0.029
50	0.050	0.021	0.030	0.025	0.035	0.021	0.029	0.024	0.030	0.021	0.030
50	0.100	0.025	0.032	0.030	0.037	0.025	0.036	0.026	0.034	0.025	0.026
50	0.200	0.021	0.023	0.032	0.032	0.031	0.033	0.031	0.028	0.025	0.025
100	0.050	0.024	0.027	0.023	0.031	0.020	0.025	0.020	0.025	0.018	0.027
100	0.100	0.020	0.024	0.023	0.027	0.015	0.027	0.016	0.026	0.018	0.028
100	0.200	0.033	0.024	0.037	0.028	0.037	0.029	0.031	0.026	0.024	0.032

Note. SD = standard deviation; LE = lower coverage error; UE = upper coverage error.

Table D 26

Lower and Upper Error of the Confidence Interval of Different Sample Sizes (True Mean = 4.4)

		Logit		Resampling		BCA		Bayesboot		Z	
n	SD	LE	UE	LE	UE	LE	UE	LE	UE	LE	UE
5	0.050	0.034	0.024	0.105	0.072	0.083	0.068	0.116	0.077	0.059	0.052
5	0.100	0.031	0.020	0.110	0.072	0.097	0.070	0.114	0.078	0.067	0.052
5	0.200	0.042	0.017	0.129	0.058	0.114	0.063	0.143	0.067	0.075	0.039
10	0.050	0.028	0.017	0.050	0.050	0.043	0.045	0.054	0.052	0.045	0.033
10	0.100	0.027	0.022	0.057	0.041	0.051	0.043	0.059	0.045	0.042	0.041
10	0.200	0.030	0.026	0.062	0.042	0.062	0.047	0.071	0.052	0.061	0.026
20	0.050	0.024	0.020	0.034	0.037	0.031	0.030	0.034	0.035	0.036	0.027
20	0.100	0.034	0.027	0.046	0.039	0.044	0.038	0.049	0.038	0.033	0.026
20	0.200	0.034	0.027	0.066	0.040	0.053	0.048	0.066	0.044	0.048	0.022
50	0.050	0.037	0.027	0.039	0.031	0.034	0.023	0.036	0.029	0.020	0.028
50	0.100	0.034	0.032	0.040	0.037	0.033	0.035	0.037	0.035	0.029	0.024
50	0.200	0.023	0.015	0.031	0.024	0.033	0.026	0.032	0.025	0.036	0.026
100	0.050	0.027	0.032	0.033	0.032	0.025	0.025	0.023	0.023	0.017	0.028
100	0.100	0.021	0.025	0.025	0.029	0.020	0.024	0.023	0.026	0.029	0.026
100	0.200	0.030	0.027	0.035	0.031	0.031	0.030	0.032	0.030	0.028	0.020

Note. SD = standard deviation; LE = lower coverage error; UE = upper coverage error.

Table D 27

Lower and Upper Error of the Confidence Interval of Different Sample Sizes (True Mean = 4.7)

		Logit		Resampling		BCA		Bayesboot		Z	
n	SD	LE	UE	LE	UE	LE	UE	LE	UE	LE	UE
5	0.050	0.039	0.014	0.114	0.062	0.090	0.063	0.115	0.075	0.063	0.044
5	0.100	0.055	0.009	0.129	0.046	0.105	0.049	0.139	0.057	0.095	0.038
5	0.200	0.062	0.012	0.153	0.041	0.129	0.037	0.161	0.048	0.113	0.018
10	0.050	0.036	0.022	0.054	0.041	0.054	0.046	0.059	0.046	0.051	0.035
10	0.100	0.037	0.020	0.080	0.042	0.070	0.042	0.082	0.043	0.067	0.027
10	0.200	0.049	0.009	0.088	0.025	0.071	0.033	0.084	0.033	0.093	0.020
20	0.050	0.035	0.026	0.052	0.035	0.046	0.037	0.050	0.037	0.037	0.032
20	0.100	0.040	0.019	0.066	0.027	0.059	0.029	0.066	0.028	0.054	0.024
20	0.200	0.044	0.018	0.058	0.027	0.043	0.032	0.053	0.030	0.065	0.014
50	0.050	0.037	0.022	0.044	0.029	0.033	0.026	0.040	0.021	0.028	0.018
50	0.100	0.022	0.029	0.027	0.033	0.023	0.037	0.025	0.035	0.041	0.025
50	0.200	0.045	0.025	0.050	0.029	0.040	0.034	0.045	0.030	0.038	0.016
100	0.050	0.025	0.031	0.027	0.034	0.021	0.030	0.023	0.028	0.031	0.030
100	0.100	0.030	0.032	0.034	0.032	0.022	0.035	0.023	0.033	0.024	0.022
100	0.200	0.033	0.016	0.031	0.018	0.024	0.023	0.025	0.016	0.028	0.016

Note. SD = standard deviation; LE = lower coverage error; UE = upper coverage error.

APPENDIX E:

Simulation Figures

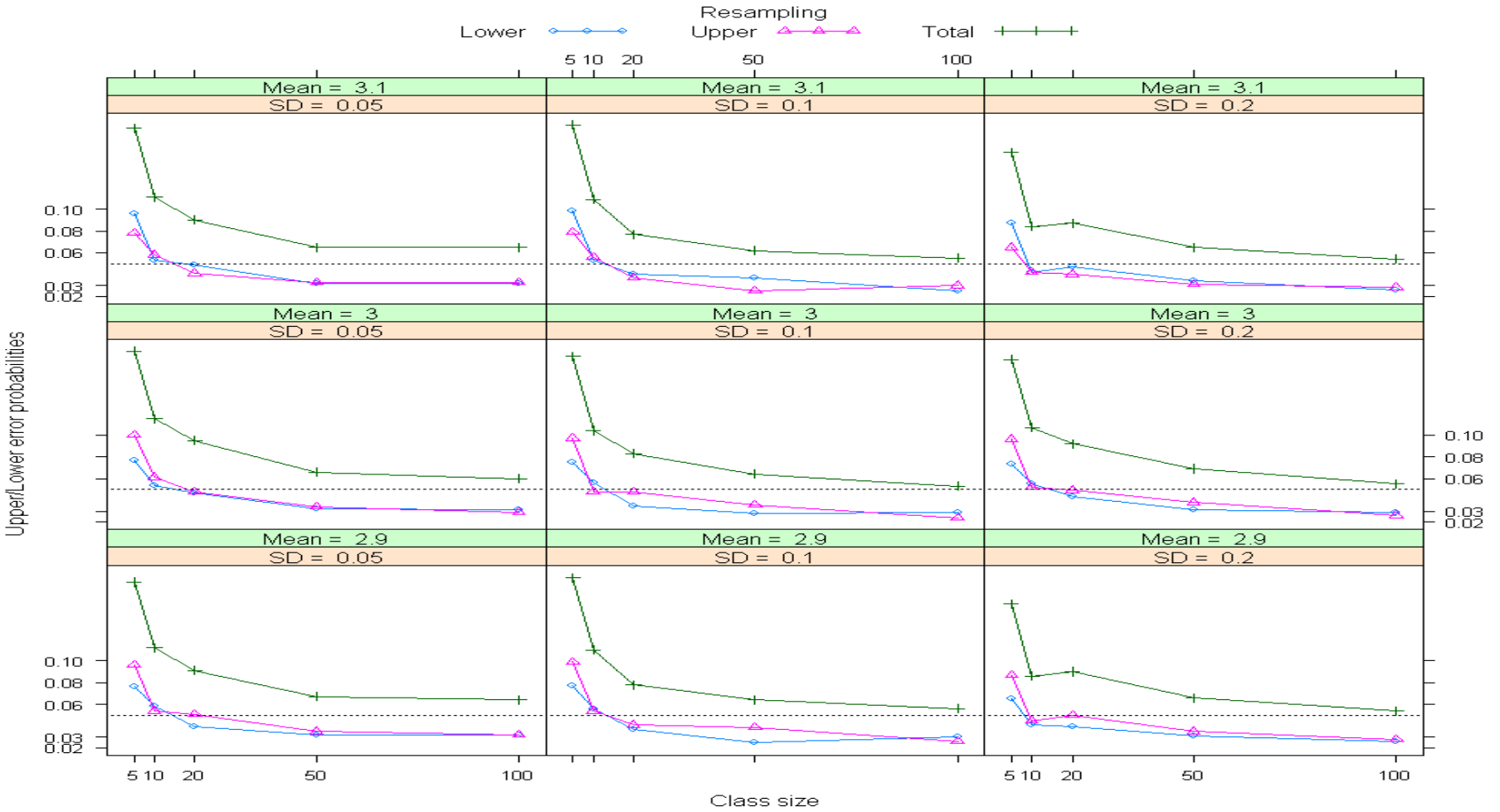

Figure E 1. Resampling Upper/Lower probabilities of the 95% CIs for the Symmetrical distributions.

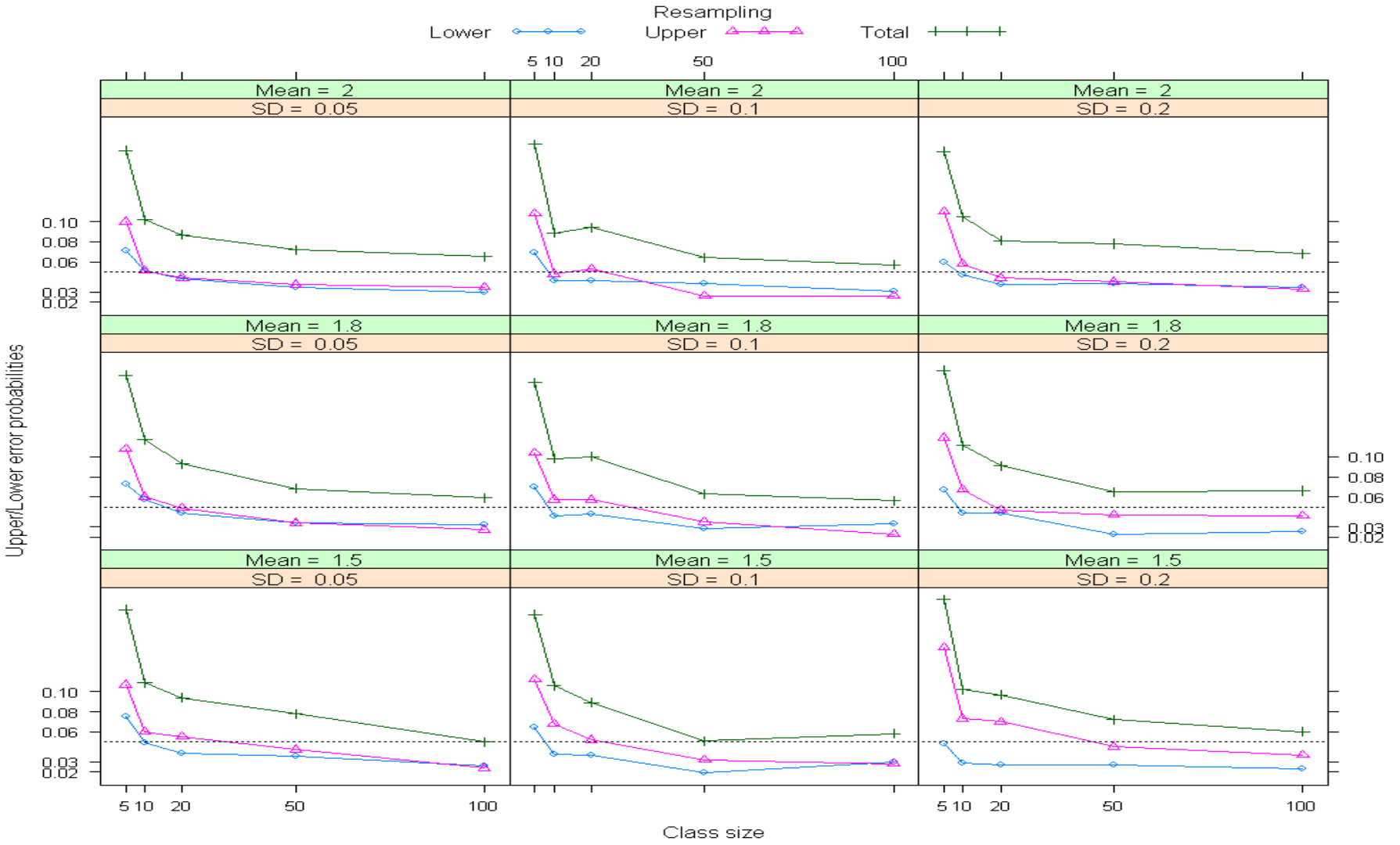

Figure E 2. Resampling Upper/Lower probabilities of the 95% CIs for the Right-skewed distributions.

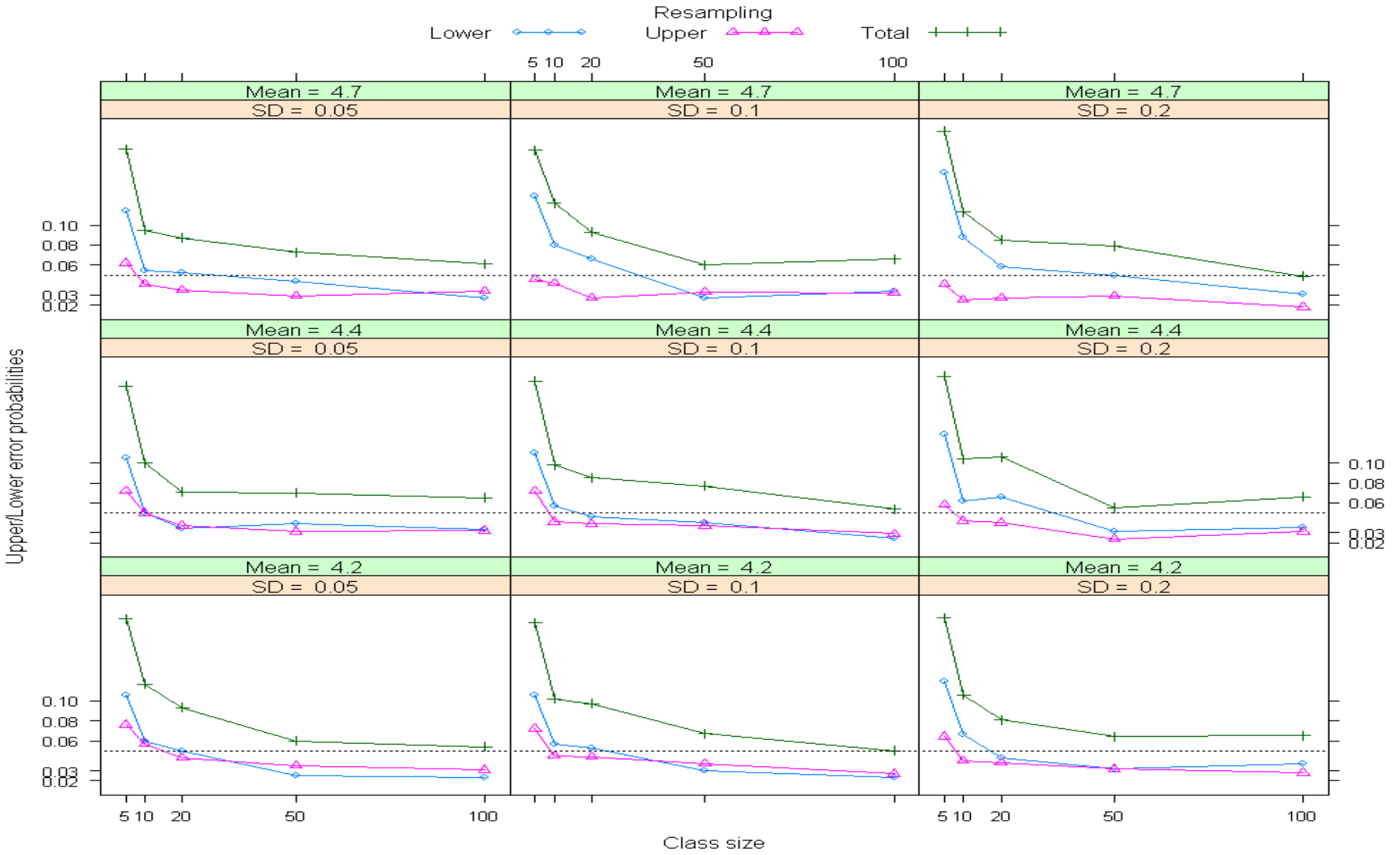

Figure E 3. Resampling Upper/Lower probabilities of the 95% CIs for the Left-skewed distributions.

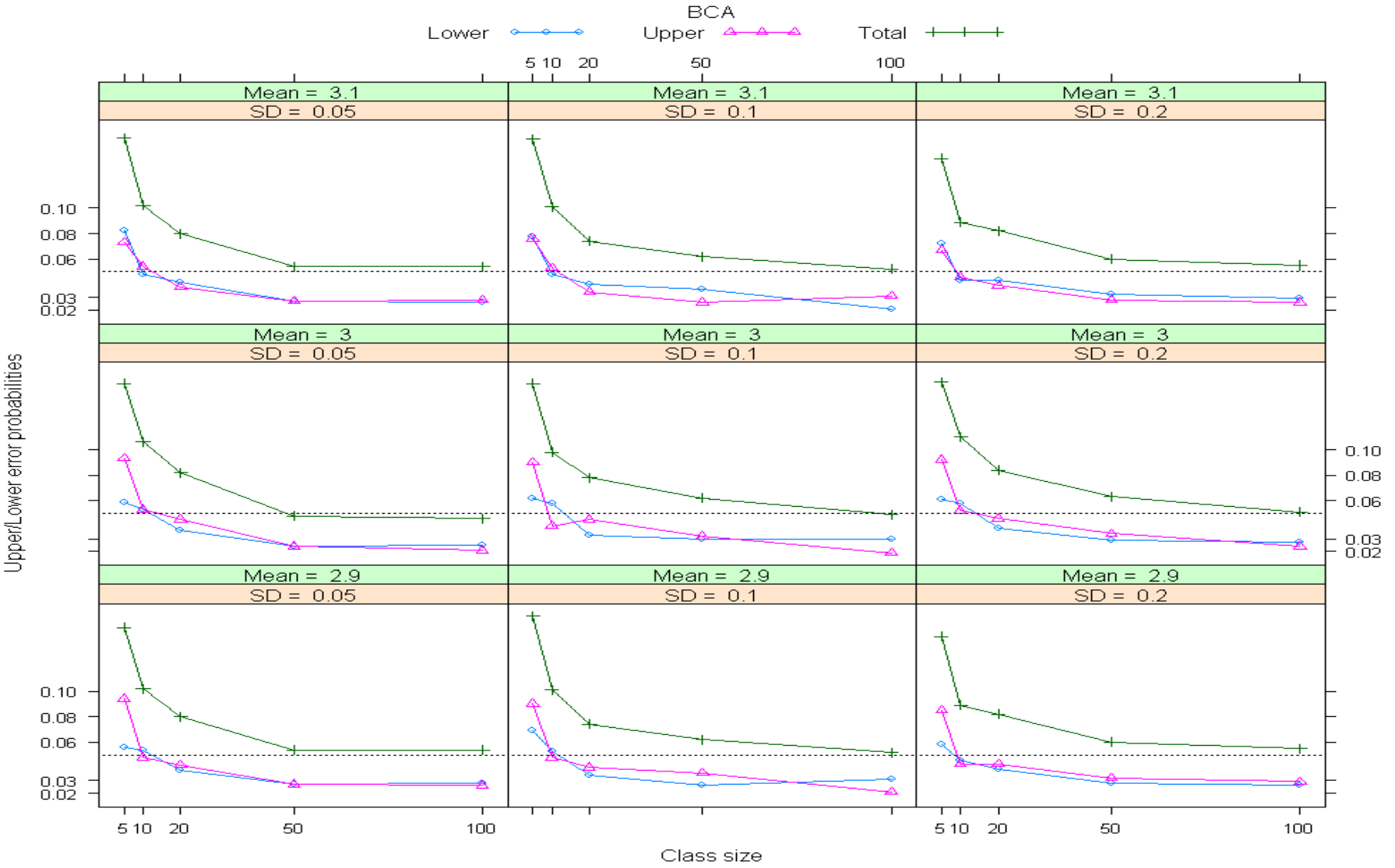

Figure E 4. BCA Upper/Lower probabilities of the 95% CIs for the Symmetrical distributions.

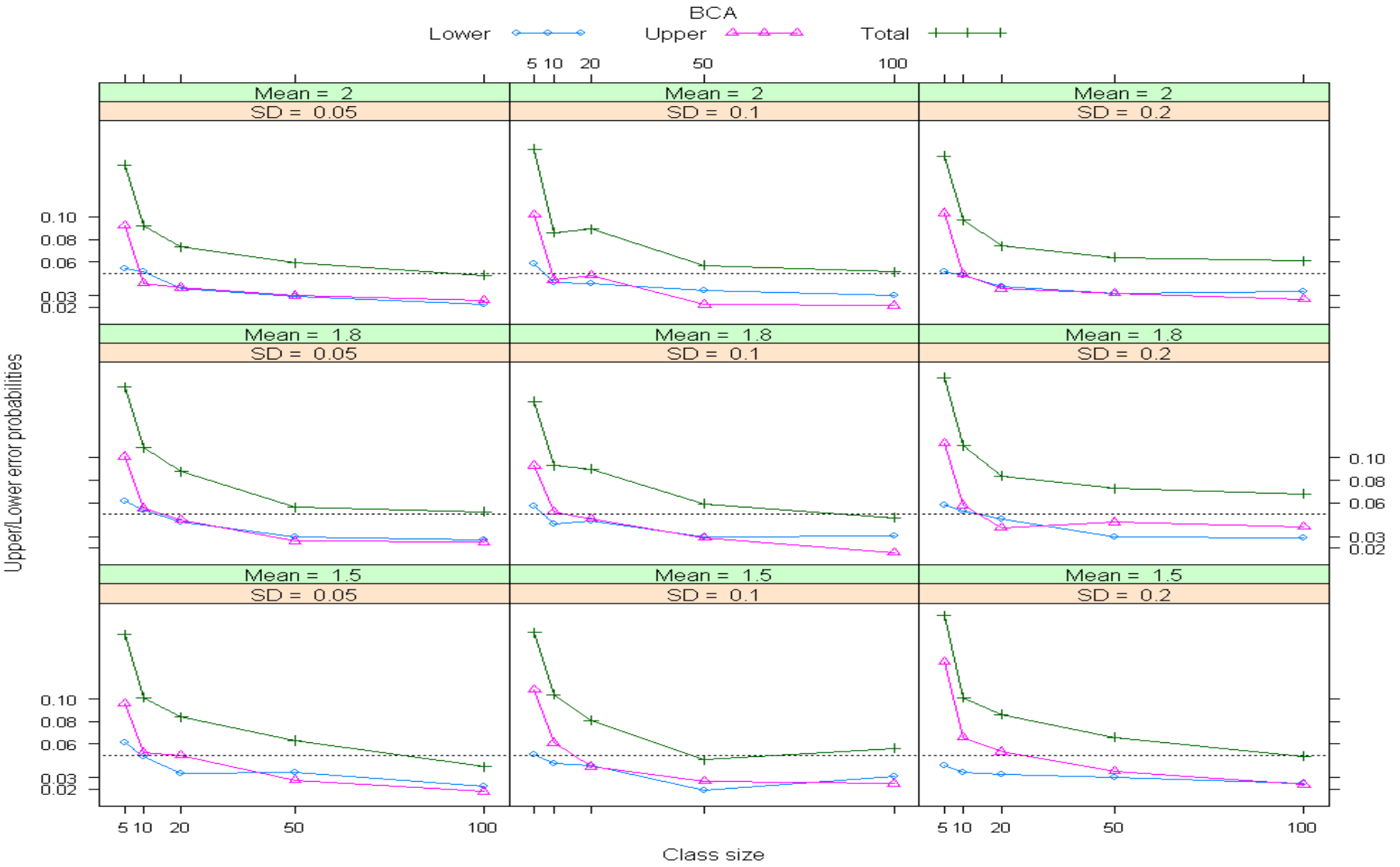

Figure E 5. BCA Upper/Lower probabilities of the 95% CIs for the Right-skewed distributions.

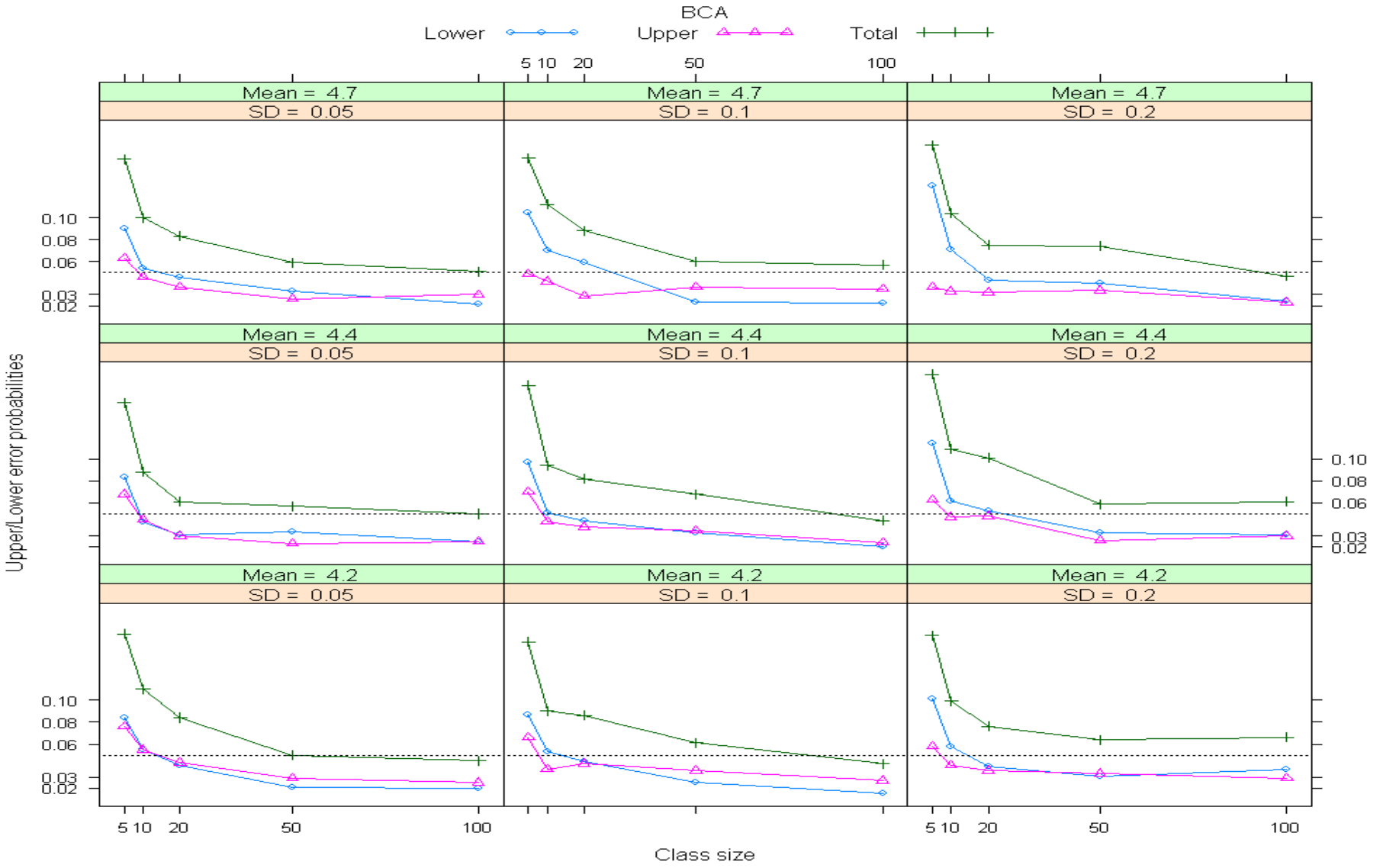

Figure E 6. BCA Upper/Lower probabilities of the 95% CIs for the Left-skewed distributions.

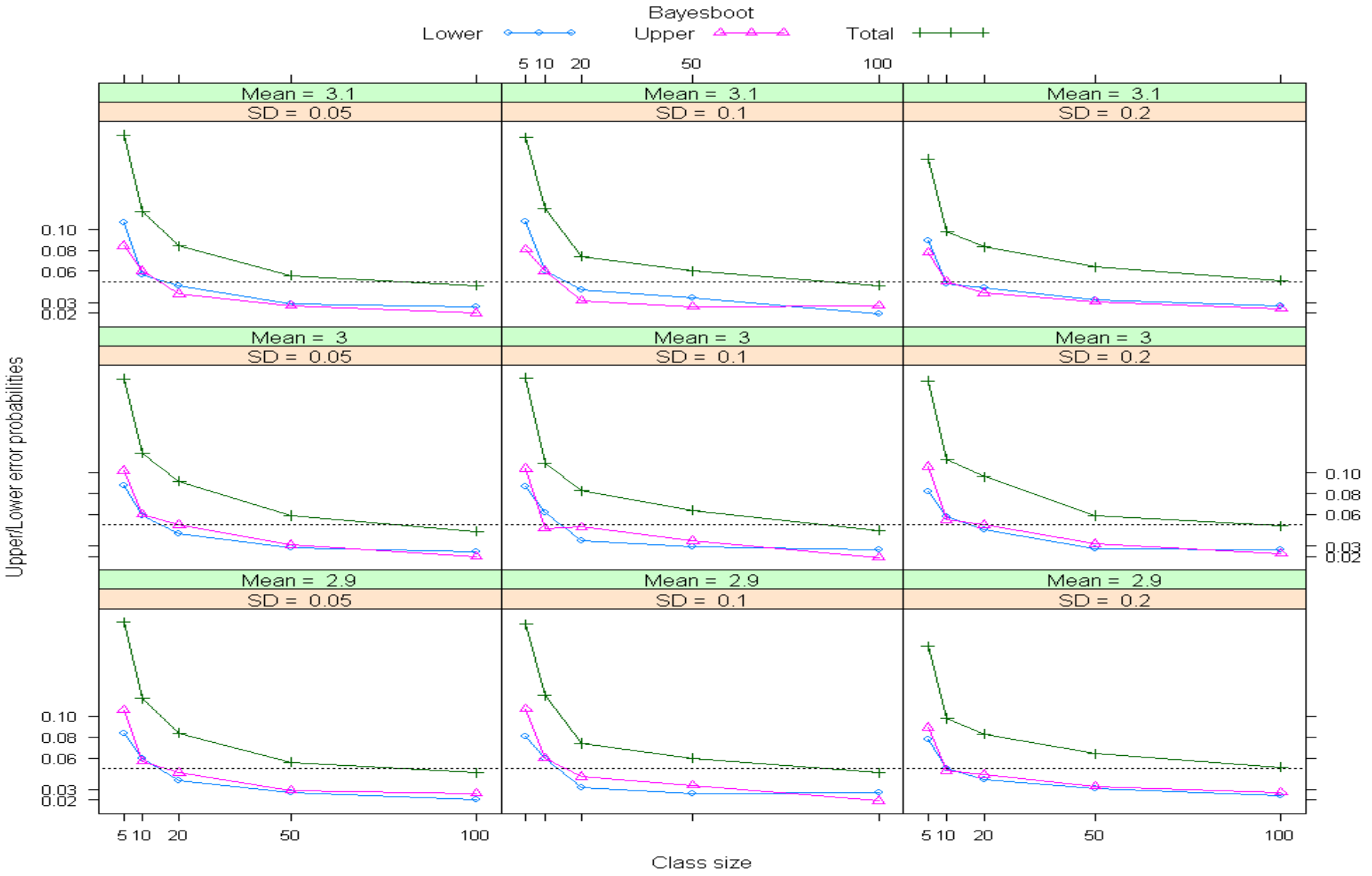

Figure E 7. Bayesboot Upper/Lower probabilities of the 95% CIs for the Symmetrical distributions.

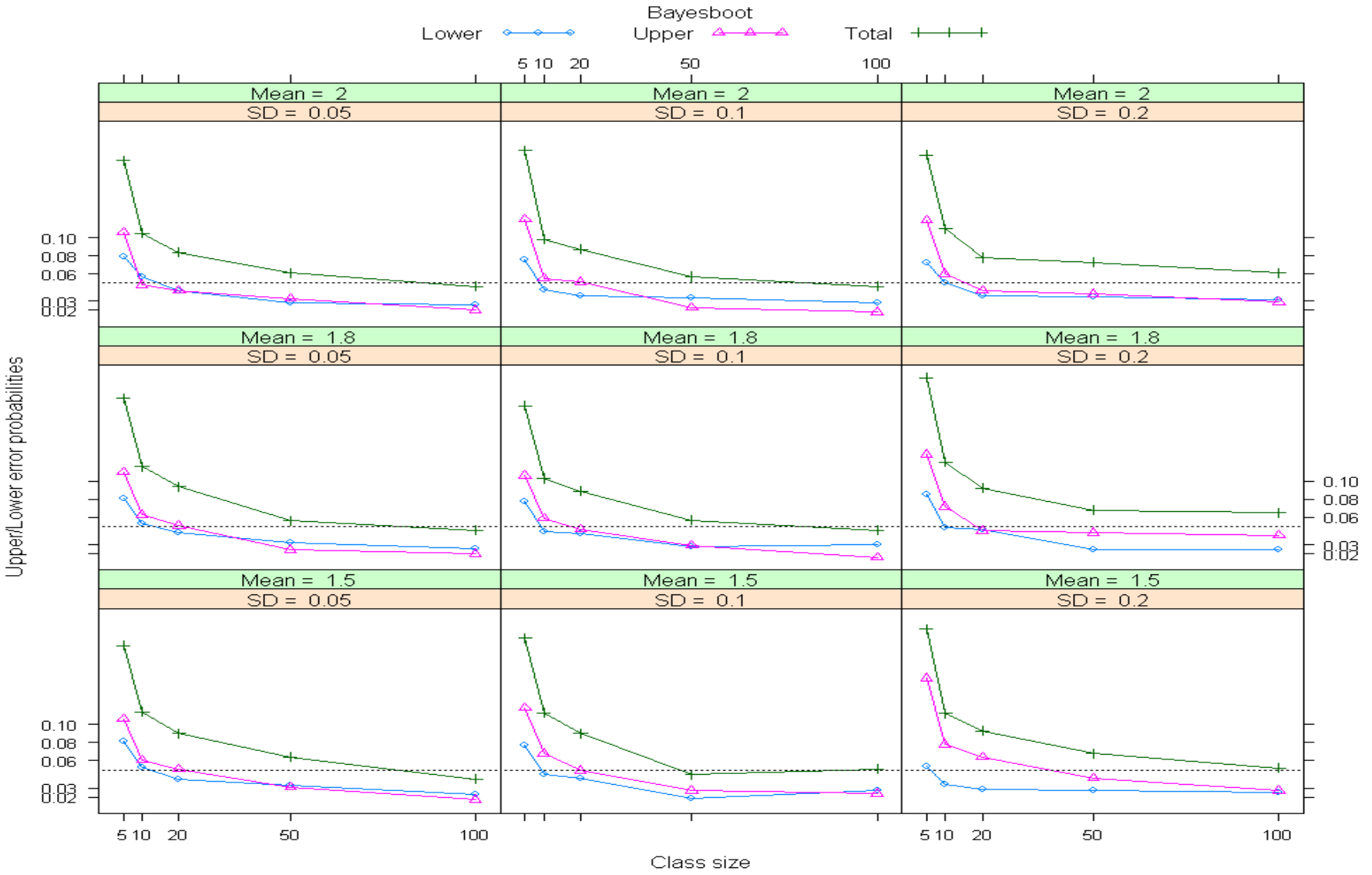

Figure E 8. Bayesboot Upper/Lower probabilities of the 95% CIs for the Right-skewed distributions.

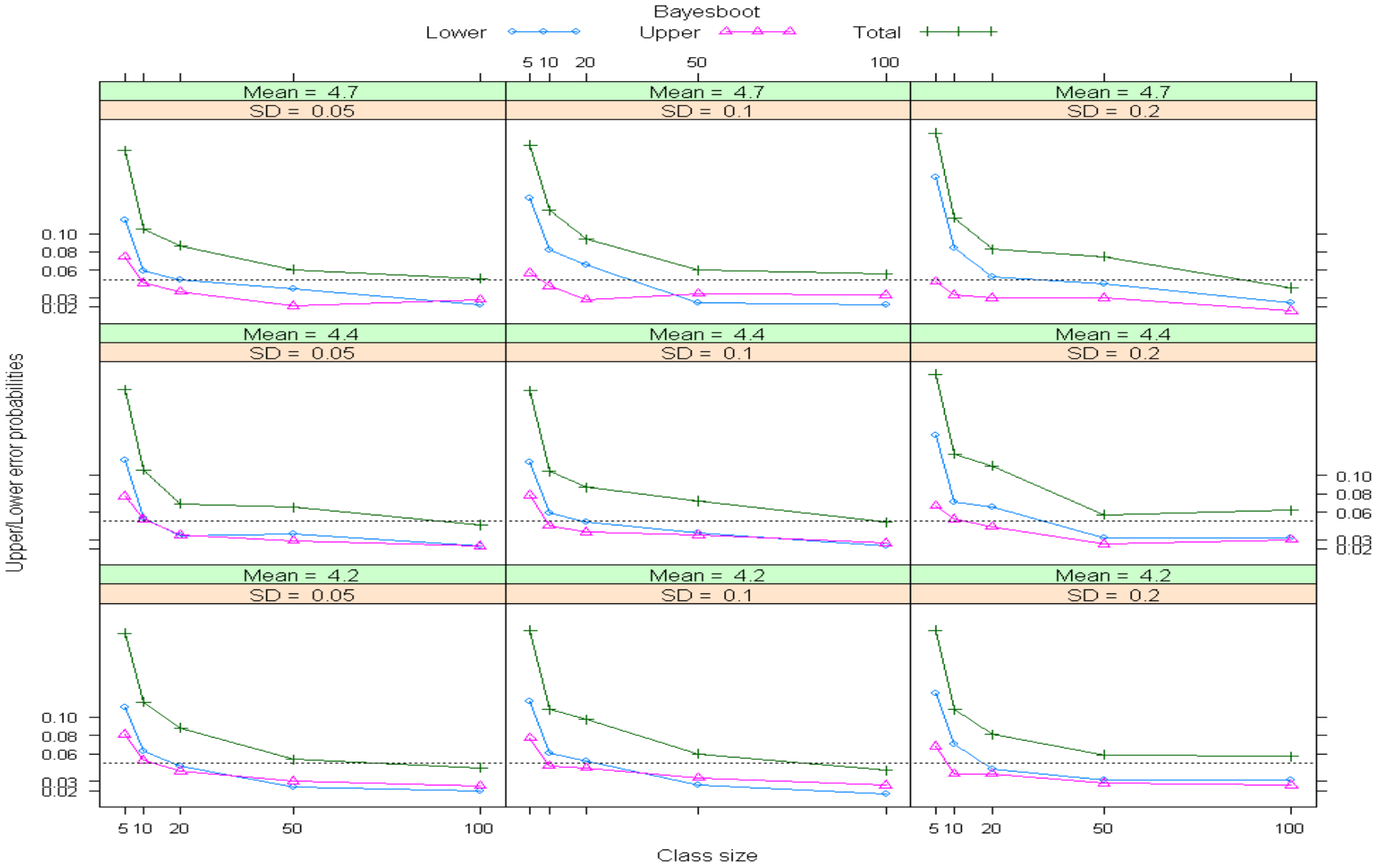

Figure E 9. Bayesboot Upper/Lower probabilities of the 95% CIs for the Left-skewed distributions.

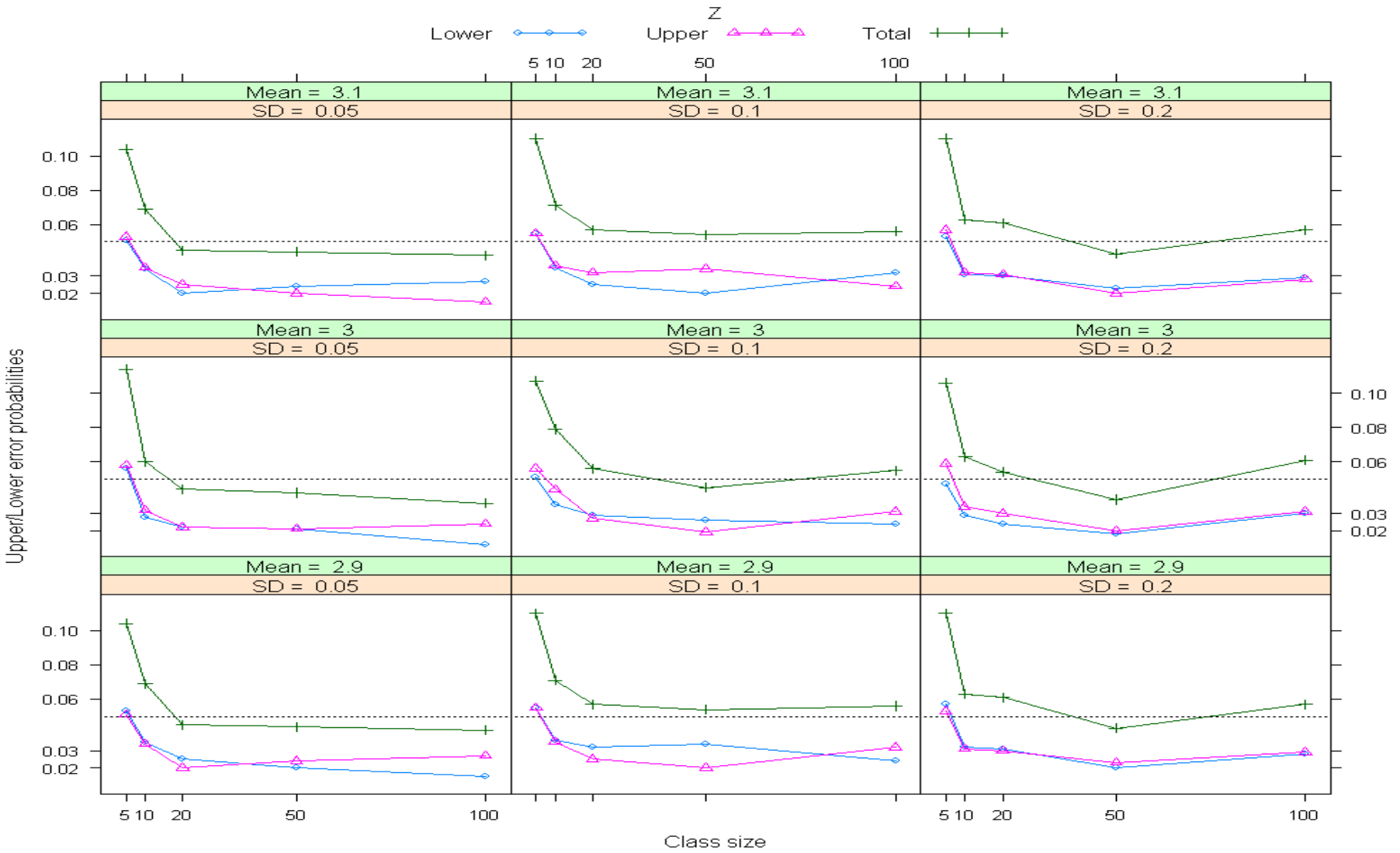

Figure E 10. Z Upper/Lower probabilities of the 95% CIs for the Symmetrical distributions.

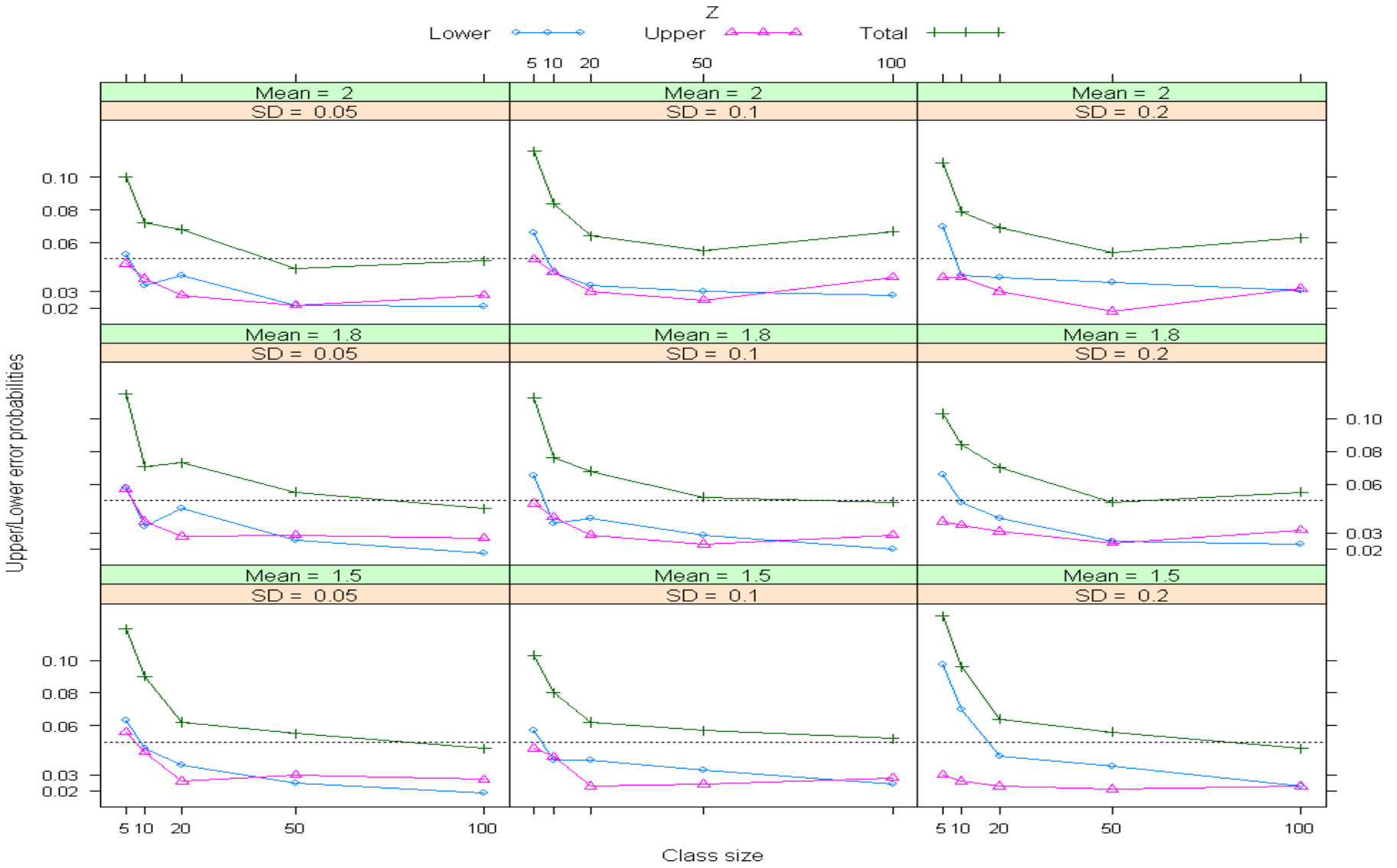

Figure E 11. Z Upper/Lower probabilities of the 95% CIs for the Right-skewed distributions.

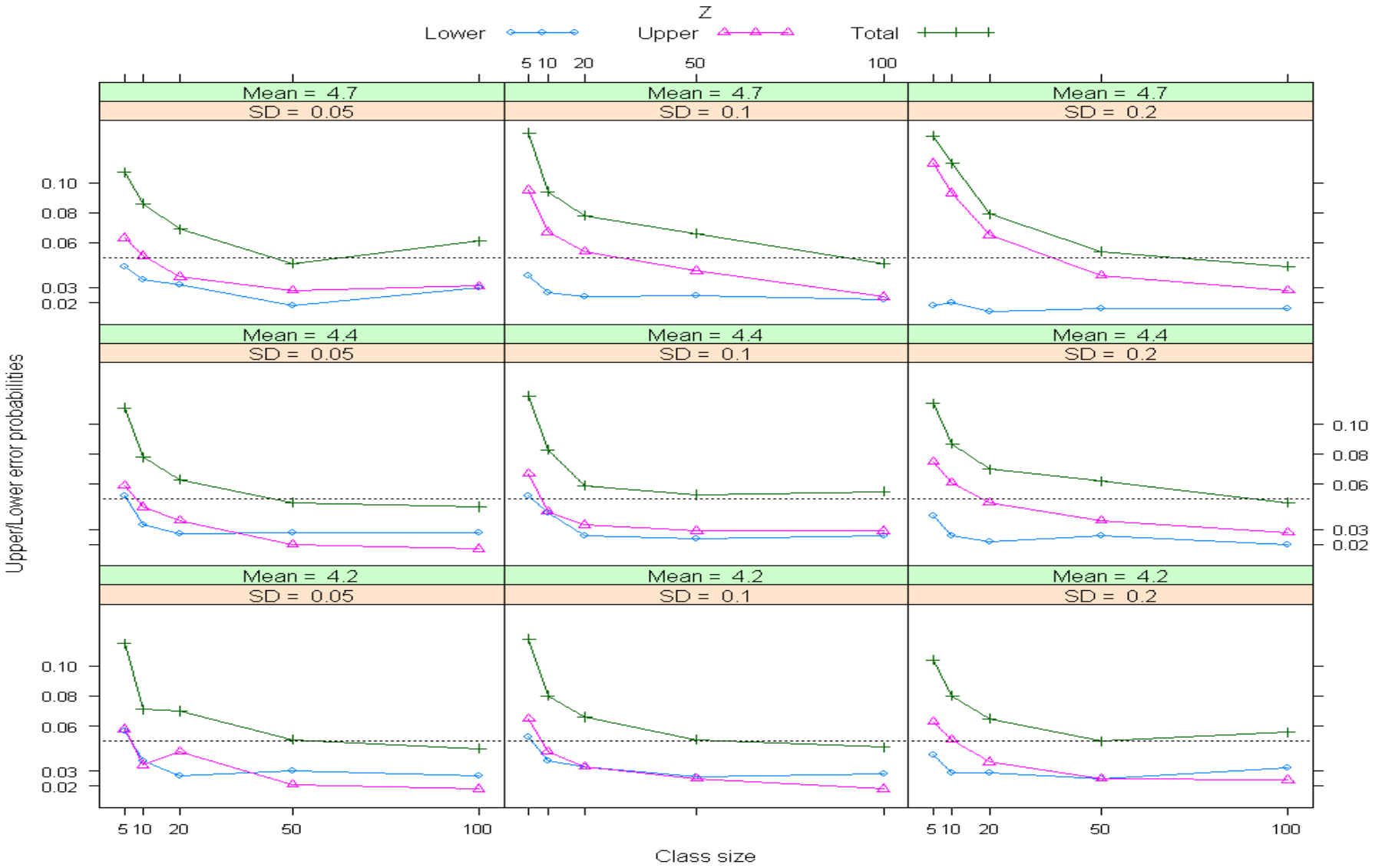

Figure E 12. Z Upper/Lower probabilities of the 95% CIs for the Left-skewed distributions.